Methods for determining the accuracy and comparing the agreement between ground measures and advanced forest inventory techniques

By Shongming Huang, Beverly Wilson, Cosmin Tansanu, Christopher W. Bater and Chao Li

This work is part of the research project on “Developing and Assessing Advanced Inventory Techniques for Enhanced Forest Management in Alberta”, supported by Canadian Forest Products Ltd. (Grande Prairie), Forsite Consultants Ltd. (Salmon Arm, British Columbia), Object-Raku Technology Inc. (Qualicum Beach, British Columbia), Forest Resource Improvement Association of Alberta (FRIAA), Canadian Wood Fibre Centre (Natural Resources Canada), ATCO Electric (Edmonton), and Alberta Agriculture, Forestry and Rural Economic Development (AFRED). Special thanks to Melonie Zaichkowsky and Dwight Weeks for their unwavering supports and contributions on many fronts under challenging circumstances, to Cam Brown, Mike Parlow and their teams and Glenn Buckmaster for their collaborations and contributions throughout this study, to Yuqing Yang and Darren Aitkin for their input and contributions at different stages, and to Lane Gelhorn for his diligent review and thoughtful questions and comments. Assistance on field data collection, including locating and felling trees, tree bucking and tree height measurements, provided by Dwight Weeks, Devin Letourneau, Clint McCrea and Karl Froese, with help from Al Benson and Tim Heemskerk, is much appreciated.

Forest Stewardship and Trade Branch

Forestry Division

Alberta Agriculture, Forestry and Rural Economic Development

Suite 303, J.G. O’Donoghue Building

7000-113 Street NW

Edmonton, Alberta, Canada T6H 5T6

Tel: 780-427-8474 or 780-422-5281

Fax: 780-427-0085

This publication is available online at https://open.alberta.ca/publications/9781460154793.

Executive Summary

Since the mid-20th century, large-scale forest inventories in many jurisdictions have relied on the stand-level interpretation of aerial photographs collected using piloted fixed-wing aircraft. Advanced forest inventory techniques from different sensors and platforms such as light detection and ranging (lidar), unmanned aerial vehicles (UAV), satellites, radar, high resolution multi-spectral and hyper-spectral imageries and other new technologies have been increasingly used in various settings to estimate forest attributes at both the area-based and individual tree scales. Two critical questions often arise in practice pertaining to the accuracy and suitability of any new forest inventory technique at the tree or stand level: (1) how can we judge whether a new inventory technique is valid or of acceptable accuracy; and (2) how can we compare if a new inventory technique agrees with or is better than an existing one. This study describes the methods that collectively provide the tools to address and answer these questions. More specifically, it focuses on the four primary inventory variables that are critically important to strategic and operational forest management, and that can be extracted directly from a potential new inventory technique: tree species, species composition, height and density. An improved error matrix is used to assess the accuracy of tree level species classification. The chi-square test and Fisher's exact test are applied to judge the similarity between species compositions from ground measures and inventory measures. Several *g*oodness-of-fit statistics and plots, an agreement measure *and* the Kolmogorov-Smirnov test are used to evaluate the accuracy and agreement (to ground measures) of height and density. The study emphasizes assessing all four primary inventory variables jointly, so that the accuracy and the level of agreement of an entire inventory technique can be judged holistically and consistently. The study also demonstrates step-by-step how to implement the methods in practice based on real world data. In addition, it also clarifies some of the basic concepts and terminologies associated with assessing forest inventory techniques, and provides the necessary technical details and caveats to interested readers on some of the issues related to error matrix, accuracy measures, statistical tests, agreement analysis and calibration.

Contents

1 Introduction

Photogrammetric and remote sensing techniques have always played an important role in operational forest inventory in Alberta. They have been used widely to gather large-scale data (i.e., area-based) to extract broad forest-level information on land stratification, wildfire, insect and disease, wildlife habitat, reforestation and climate change. With the rapid development of the new technologies in terrestrial and airborne light detection and ranging (lidar, LiDAR or LIDAR), unmanned aerial vehicles (UAV, also known as remotely piloted aircraft systems (RPAS), drones), satellites, radar, digital aerial photogrammetry (DAP), and high resolution multi-spectral and hyper-spectral imageries, they have also been increasingly used to collect fine-scale data at the tree level, to identify individual tree stems (crowns) and tree species in ways and spatial resolutions previously unseen. There is a real hope and optimism that, if proven to be accurate and more effective, efficient (i.e., reduced costs at increased speed of inventory), reliable, consistent and operational, some of the new aerial-based remote sensing techniques or other emerging state-of-the-art technologies can be transformative and become the new paradigm for enhanced forest inventory (EFI) and for accurate and timely forest population census.

Two critical questions arise in practice with regard to the use of any new technique in forest inventory:

- How can we judge whether a new inventory technique is valid or good enough?
- How can we compare if a new inventory technique agrees with or is better than an existing one?

For instance, we often ask: is the new inventory technique (e.g., from lidar, UAV or satellite) good to be implemented in practice at the stand level and/or tree level? Is the new lidar, UAV, satellite or any other new technique better than the high spatial resolution stereo imagery in a softcopy environment currently used in the Alberta Vegetation Inventory (AVI)? Do the new and old/existing techniques agree with each other in getting the primary forest inventory variables? Can the new technique be semi-automated or fully automated for forest population census? How well do all these techniques agree with ground/field observations? Can the new technique be used to supplant the ground measurement of permanent sample plots (PSPs) and temporary sample plots (TSPs)?

This study is designed to present the statistical methodologies capable of answering the above questions and judging the competing new inventory techniques consistently. While the new inventory techniques may have many great features and promising new capabilities, practitioners should not automatically assume that they would work in operations without assessing their accuracy and agreement relative to ground observations. We have seen some new and improved techniques that appear very impressive within the research realm, and on paper and in graphics, only to fail miserably when used in operations in the real world. This is very relevant in photogrammetric and remote sensing studies, as some of the new techniques may still not meet the operational accuracies and resolutions at the tree or stand level, or require excessive and unrealistic amount of resources (time, cost and expertise) to implement on a large population scale. Therefore, they may still only meet the coarse, broad level forest information needs perhaps for strategic purposes, rather than the operational forest inventory requirements at the tree level for forest management and operational planning.

Since “forest inventory” is a wide-ranging concept that can include, among others, land stratification, forest cover type specification, sampling, and forest health, forest hydrography, forest fire hazard, topography, ecosystem, biodiversity, stand structure, demography, insect and disease, climate change and habitat assessments, our focus is narrowed to the most common mensurational forest inventory that involves only tree level variables and stand level variables summarized from the tree level variables. More specifically, our focus is on four primary inventory variables that are critically important to operational forest management, and that can be extracted directly from promising new inventory techniques at the tree level:

- Species, species composition, species-specific height and density.

These four primary inventory variables are considered the fundamental base variables in any tree or stand level forest inventory. The method comparison methodologies described in this study would also apply to assessment of many other tree, stand and landscape level variables, such as: diameter at breast height (DBH), DBH distribution, basal area, quadratic mean diameter, site index, age, total and merchantable volumes, defect, piece size, log profile, natural and anthropogenic mortality, ingress, biomass and carbon. However, these attributes are typically derived indirectly through other additional processes, and as such, they can be highly impacted by the quality of the primary variables and other indirect and non-inventory factors incurred in the processes, including the choices of the auxiliary models, the modeling approaches taken, the viability and strength of the connecting relationships implicated, and the calibration techniques chosen. In some cases, like for volume, biomass or carbon, this may be further complicated by the ancillary data and “helper” models used to estimate volume, biomass or carbon from tree level variables. The goodness of the estimation may not be indicative of the accuracy and

agreement of an inventory technique (it may indicate the quality of other data, models or processes involved). For simplicity of illustration, we have focused our discussion on attributes directly extracted from detailed tree level inventory techniques.

In order to be considered a better forest inventory technique, among many factors that need to be considered (such as time, cost or economic viability, relevant expertise and required resources, and other operational objectives, variables and constraints), the inventory technique must demonstrate, at a minimum, the accuracy and agreement to ground measures in:

1). Identifying individual tree species (species);

2). Representing the forest-level species frequency distribution (species composition);

3). Characterizing the values and the distribution of height (height);

4). Segmenting individual stems (density).

Technically "segmenting individual stems" (stem segmentation) is "identifying individual crowns" or crown delineation. In this study, we use the terms interchangeably, although crown delineation is a broader concept that can also involve delineating the size and shape of crowns, besides just identifying the existence of the crowns.

The main objectives of this study are to:

1. Describe the methods that can be used consistently to determine the accuracy of forest inventory techniques and that can compare the agreement between ground measures and forest inventory techniques;

2. Demonstrate step-by-step how to implement the methods in practice based on the real world data and examples;

3. Clarify some of the basic concepts and terminologies associated with assessing forest inventory techniques, and provide the necessary technical details, justifications and caveats to interested readers on some of the issues related to error matrix, accuracy assessment for categorical and continuous variables, statistical test, agreement analysis and calibration.

To achieve the stated objectives, the methods for determining the accuracy of species classification are described first in Section 2, followed by the methods for assessing species composition (i.e., species frequency distribution) in Section 3. Section 4 presents the recommended statistics and methods for assessing two other primary forest inventory variables, height and density. For each method an example based on real world data is given to demonstrate how to implement the method and interpret the results in practice. Section 5 provides additional explanations, technical details and cautionary notes on some of the issues related to the recommended methods and some other relevant methods. It essentially answers the "whys" on many topics touched on in this study. It also points out some frequent confusions and misunderstandings in some previous accuracy assessment studies and offers suggestions on how to use and interpret the recommended methods correctly. Finally Section 6 summarizes the results of this study and presents concluding remarks and recommendations.

Most of the previous studies focused on the accuracy assessment for up to three primary inventory variables (e.g., species, height and/or density). This study focuses on the complete methods that address all four primary inventory variables, so the validity and the level of agreement (to ground measures) of an entire new inventory technique can be judged holistically and consistently.

This study does not assess the explicit remote sensing technologies, algorithms and metrics used to derive the inventory variables, as other researchers and specialists are better qualified for this. It is not our objective to look into, for example, how to extract tree height, predict tree species, delineate tree crowns or make stem segmentation from lidar point clouds or lidar metrics; how to delineate stands or polygons using image segmentation algorithms and generate stand structure from pixel-based image compositing; or how to predict forest metrics/attributes using artificial intelligence, machine learning or the random forests method based on the K-nearest neighbor imputation algorithms. Instead, we are interested in looking at the variables resulting from an inventory technique, however they are extracted, through whatever technologies, algorithms, approaches or methods.

To allow interested readers to verify or duplicate all computations and results involved in this study and avoid the difficulties of not knowing the exact data, all data sets used in this study are listed in relevant tables. Many are also displayed in graphical forms. Since this study is mainly intended to be methodological, i.e., to describe the fundamental concepts, approaches and methodologies, rather than an explicit evaluation of a specific forest inventory technique, the sample sizes of the selected data sets are not large. However, it should not be difficult to conceive that the concepts, approaches and methodologies discussed and demonstrated in this study can be applied to relevant categorical and continuous variables to assess the accuracy and

agreement of any forest inventory techniques. It is our hope that the concepts and methodologies presented in this study can become an invaluable source for those who are interested in the fundamentals, and most importantly, practical solutions for determining the accuracy of forest inventory techniques and comparing the agreement between ground measures and forest inventory techniques.

It is worth noting ahead that some of the concepts and approaches presented in this study are different from the conventional ones appeared in previous research or used elsewhere, and several of which are intrinsically complex and multifaceted. As a result, it could be difficult to comprehend the full contents of this study without some protracted efforts and additional experience and background. We have organized the document to highlight the basic concepts and step-by-step examples in Sections 2 to 4, with some of the more detailed technical and supplementary materials provided in Section 5 for interested readers.

Most of the concepts and approaches presented in this study are not new – they have appeared in other scientific disciplines, but they are woven together in a holistic manner in this study. Readers are invited to rethink and move towards an integrated holistic approach of assessing forest inventory techniques, not just merely focusing on a single variable, a single case, or a solitary statistic or statistical test.

One other point worth noting ahead is that some researchers may find that some of the analyses and discussions presented in this study may appear critical and opinionated in some regards, as we aim to point out the misunderstandings, misuses, limitations and deficiencies of some of the conventional concepts and approaches in previous research. For practical reasons we want to be very clear and precise on the concepts and approaches we advocate and why. From the start, it was decided that we were not interested in research for research's sake, but wanted to see research that can be put into practical real world use, not just in academic research realms. It was also decided early on to show exactly what we did. All the data and formulas used in the analyses are provided for interested readers to check and verify. We will leave them up to the readers to ponder, judge and decide for themselves. However, if anyone who thinks that we have been too critical of his/her published work, bear in mind that we are only criticizing a small part of the published work. We have learned and benefited from other parts of the published work. The criticisms are intended to clarify the possible misconceptions, avoid repeating the prevailing biases and mistakes, improve understanding and advance knowledge in assessing forest inventory techniques. In spite of the striving efforts, our presentation is not immune to its own limitations and shortcomings. We are certain that there are different viewpoints, objectives, focuses and preferences. We welcome and appreciate any comments or counter-criticisms about our work.

2 Accuracy of tree level species classification

The accuracy of tree level species classification from an inventory technique is judged by the correctly classified proportions or percentages for the species, through an error matrix, or the species classification performance matrix, which is a special case of the error matrix applied to the categorical variable "species".

For reference in the examples, the main tree species in Alberta are listed in Table 1, alongside the matching species code, group, type and scientific name for each species. The scientific names for dead pine and snag/dead tree stump are direct Latin translations. Dead pine is used to denote dead pine trees killed by mountain pine beetles.

TABLE 1. LIST OF MAIN TREE SPECIES IN ALBERTA.

Type	Common name	Scientific name	Code	Group
Deciduous	Aspen	*Populus tremuloides* Michx.	Aw	AwPb
	Balsam poplar	*Populus balsamifera* L.	Pb	AwPb
	White birch	*Betula papyrifera* Marsh.	Bw	Bw
Coniferous	Lodgepole pine	*Pinus contorta* var. *latifolia* Engelm.	Pl	Pine
	Whitebark pine	*Pinus albicaulis* Engelm.	Pw	Pine
	Limber pine	*Pinus flexilis* E. James	Pf	Pine
	Jack pine	*Pinus banksiana* Lamb.	Pj	Pine
	White spruce	*Picea glauca* (Moench) Voss	Sw	SwFir
	Engelmann spruce	*Picea engelmannii* Parry ex Engelm.	Se	SwFir
	Balsam fir	*Abies balsamea* (L.) Mill.	Fb	SwFir
	Alpine fir	*Abies lasiocarpa* (Hook.) Nutt.	Fa	SwFir
	Douglas-fir	*Pseudotsuga menziesii* (Mirb.) Franco	Fd	SwFir
	Black spruce	*Picea mariana* (Mill.) B.S.P.	Sb	SbLt
	Tamarack larch	*Larix laricina* (Du Roi) K. Koch	Lt	SbLt
	Western larch	*Larix occidentalis* Nutt.	Lw	SbLt
	Alpine larch	*Larix lyallii* Parlatore	La	SbLt
	Dead pine*	*Mortuus abiete* (Latin)	Dp	Dp
Coniferous/deciduous	Snag/dead tree stump	*Arboris truncus mortuus est* (Latin)	Sg	Snag

Note: *Dead pine is used to denote dead pine trees killed by mountain pine beetles.

2.1 Error matrix for species classification

For species classification, the error matrix with k possible species (i.e., k categories) is a $k \times k$ contingency table (also known as a two-way table, a bivariate table, a square table, or a crosstab) of frequency counts by category. It is shown in Table 2, where each cell n_{ij} in the table is the count that corresponds to the category (i=1, 2, ⋯, k; j=1, 2, ⋯, k; or simply, i, j=1, 2, ⋯, k), and N is the total number of counts. All other variables are defined in detail below.

The error matrix in Table 2 is integrated with four accuracy proportions (to be discussed next) and can also be integrated with other quantities (to be discussed later), hence it can be termed an integrated error matrix. Although not critical and does not change the values, the integrated error matrix lists the "classification" from an aerial-based remote sensing technique on top (as columns, since it is typically described from above), and the "reference" on level (as rows, since ground is typically used as the reference). This arrangement is different from many other error matrices where "reference" often appears on top and "classification" often appears on level. The intricacy of this seemingly trivial switch between row and column of the error matrix should become clear later.

TABLE 2. AN ERROR MATRIX WITH *K* CATEGORIES (SPECIES) AND *N* TOTAL NUMBER OF COUNTS.

		Classification (inventory)				Total	Correct proportion re
	Category	1	2	⋯	k	(row)	reference
Reference (ground)	1	n_{11}	n_{12}	⋯	n_{1k}	nr_1	$PR_1 = n_{11}/nr_1$
	2	n_{21}	n_{22}	⋯	n_{2k}	nr_2	$PR_2 = n_{22}/nr_2$
	⋮	⋮	⋮	⋱	⋮	⋮	⋮
	K	n_{k1}	n_{k2}	⋯	n_{kk}	nr_k	$PR_k = n_{kk}/nr_k$
	Total (column)	nc_1	nc_2	⋯	nc_k	N	
	Correct proportion re classification	$PC_1 = n_{11}/nc_1$	$PC_2 = n_{22}/nc_2$	⋯	$PC_k = n_{kk}/nc_k$		$P_o = \frac{\sum_{i=1}^{k} n_{ii}}{N}$
	Average of correct proportions by category	$PAve_1$	$PAve_2$	⋯	$PAve_k$		

Note: detailed definitions for the variables appeared in the table are provided in the main text.

In the integrated error matrix (Table 2), the reference counts are represented by the rows and the classifications are represented by the columns. The shaded values (n_{ii}s) in the diagonal represent the correctly classified counts by the classification for the categories. More specific definitions and explanations for the terms and variables appeared in Table 2 are provided here:

Reference – typically refers to the ground measures that are considered to represent the truth or "ground truth". It can also denote any reference measures that are considered the commonly accepted "gold" standards, or the consensus (e.g., sometimes the interpreted AVI data are used as the reference standard against the other types of data in comparison and calibration).

Classification – refers to the classifications, interpretations, observations, measures, calls, estimates or predictions from aerial methods or remote sensing techniques. It can also denote the classification obtained from a "map", an "image", or an "inventory" derived from an inventory technique. Therefore, these terms, "classification", "map", "image", "inventory", "inventory classification", "inventory measure", "prediction" and "estimate" are sometimes used interchangeably in this study.

Row, column and grand totals – row totals are denoted by nr_i and column totals are denoted by nc_i. For instance, nr_1, nr_2, ⋯, nr_k denote row totals for rows 1, 2, ⋯, k, respectively; nc_1, nc_2, ⋯, nc_k denote column totals for columns 1, 2, ⋯, k, respectively. For the kth category, the row total $nr_k=n_{k1}+n_{k2}+\cdots+n_{kk}$ and the column total $nc_k=n_{1k}+n_{2k}+\cdots+n_{kk}$. The grand total ($N$) is the sum of row totals $\sum_{i=1}^{k} nr_i$ or column totals $\sum_{i=1}^{k} nc_i$ (recognizing i, j=1, 2, ⋯, k).

Important Accuracy Measures from the Error Matrix

Four important accuracy measures (also known as accuracy proportions or accuracy percentages) are calculated and included in the integrated error matrix (Table 2):

1. **Overall accuracy** (P_o) – refers to the overall classification accuracy proportion or percentage when all categories are combined. It is calculated by dividing the sum of the shaded diagonal values (correct counts for the categories) in the error matrix (Table 2) by the total number of observations (N):

   [2.1] $$P_o = \frac{n_{11}+n_{22}+\cdots+n_{kk}}{N} = \frac{\sum_{i=1}^{k} n_{ii}}{N} \qquad (P_o\% = 100 \times \frac{\sum_{i=1}^{k} n_{ii}}{N}).$$

   The overall accuracy represents one of the most important and commonly cited measures that characterize the overall accuracy of a classification when all categories in the classification are combined.

   Besides the overall accuracy, it is also very useful to know how this overall accuracy is distributed across different individual categories. Due to their biological and morphological characteristics and the unique inventory technique that may or may not favor some categories (i.e., species) in certain growing conditions, seasons (e.g., leaf-on, leaf-off) and structures, the categories in the inventory could have very different accuracies than that of the overall accuracy. Hence, the accuracies for individual categories are often needed in order to fully assess the accuracies from an inventory technique. For a k×k table the overall accuracy can be distributed to each category by row, by column, and by row and column combined.

2. **PR (correct Proportion or Percent relative to the Reference)** – refers to the correct proportion (or percentage, i.e., % correct) of a category when compared to the total number of counts from the reference for that category. For example, what percent of the white spruce trees in the ground sample were correctly classified as white spruce in the inventory? It is calculated by dividing the number of correctly classified counts for the category, by the total number of counts from the reference for that category (row total). In a general form it can be written for category i as:

   [2.2] $$PR_i = \frac{n_{ii}}{nr_i} \qquad (PR_i\% = 100 \times \frac{n_{ii}}{nr_i}).$$

   where n_{ii} is the number of correctly classified counts for category i, and nr_i (row total) is the total number of counts from the reference for category i (i=1, 2, ⋯, k). The pooled or weighted average of PR_i is $(PR_1\times nr_1+PR_2\times nr_2+\cdots+PR_k\times nr_k)/(nr_1+nr_2+\cdots+nr_k) = \sum_{i=1}^{k}(PR_i \times nr_i) / \sum_{i=1}^{k} nr_i = \sum_{i=1}^{k} n_{ii} / N = P_o$, which means that the overall accuracy P_o can be distributed by row.

   The PR has traditionally been referred to as "producer's accuracy". We will explain why we prefer "PR" over "producer's accuracy" or any other term later.

3. **PC (correct Proportion or Percent relative to the Classification)** – refers to the correct proportion (or percentage, i.e., % correct) of a category when compared to the total number of counts classified as that category by the inventory

classification. For example, what percent of the trees classified as white spruce in the inventory are actual white spruce trees on the ground? It is calculated by dividing the number of correctly classified counts for the category, by the total number of counts classified as that category by the classification (column total). In a general form it can be written for category *i* as:

[2.3] $\mathrm{PC}_i = \frac{n_{ii}}{nc_i}$ $(\mathrm{PC}_i\% = 100 \times \frac{n_{ii}}{nc_i})$.

where n_{ii} is the number of correctly classified counts for category *i*, and nc_i (column total) is the total number of counts classified as category *i* by the classification. The pooled average of PC_i is $(\mathrm{PC}_1 \times nc_1 + \mathrm{PC}_2 \times nc_2 + \cdots + \mathrm{PC}_k \times nc_k)/(nc_1 + nc_2 + \cdots + nc_k) = \sum_{i=1}^{k}(\mathrm{PC}_i \times nc_i) / \sum_{i=1}^{k} nc_i = \sum_{i=1}^{k} n_{ii} / N = \mathrm{P_o}$, which means that the overall accuracy $\mathrm{P_o}$ can also be distributed by column.

The PC has traditionally been referred to as "user's accuracy" or "reliability". We will explain why we prefer "PC" over "user's accuracy" or "reliability" or any other term later.

4. **PAve (Average of correct Proportions for category)** – refers to the pooled or weighted average of PR and PC. It can be written in the following general form for category *i*:

[2.4] $\mathrm{PAve}_i = \frac{\mathrm{PR}_i \times nr_i + \mathrm{PC}_i \times nc_i}{nr_i + nc_i} = \frac{(n_{ii}/nr_i) \times nr_i + (n_{ii}/nc_i) \times nc_i}{nr_i + nc_i} = \frac{2n_{ii}}{nr_i + nc_i}$

$(\mathrm{PAve}_i\% = 100 \times \frac{2n_{ii}}{nr_i + nc_i})$.

More intuitively, the average of correct proportions for a category can be written as:

[2.5] $\mathrm{PAve}_i = \frac{row\ correct\ count + column\ correct\ count}{row\ total + column\ total}$.

where the row and column correct counts and the row and column totals correspond to category *i*. It can be shown that the pooled average of PAve_i is also identical to the overall accuracy $\mathrm{P_o}$ (i.e., the overall accuracy can be distributed into PAve_i):

[2.6] $\mathrm{P_o} = \frac{\mathrm{PAve}_1(nr_1 + nc_1) + \mathrm{PAve}_2(nr_2 + nc_2) + \cdots + \mathrm{PAve}_k(nr_k + nc_k)}{(nr_1 + nc_1) + (nr_2 + nc_2) + \cdots + (nr_k + nc_k)} = \frac{\sum_{i=1}^{k}[\mathrm{PAve}_i(nr_i + nc_i)]}{\sum_{i=1}^{k}(nr_i + nc_i)} = \frac{\sum_{i=1}^{k} 2n_{ii}}{N+N} = \frac{\sum_{i=1}^{k} n_{ii}}{N}$.

The four accuracy proportions ($\mathrm{P_o}$, PR, PC and PAve) calculated from the error matrix are used as the four accuracy measures for species classification. They represent overall ($\mathrm{P_o}$) and individual species level (PR, PC and PAve) classification accuracies. There are numerous (40+) other accuracies and error measures that can be derived from an error matrix, but they are not as useful nor as meaningful in practical accuracy assessment as the four measures described above. Interested readers may wish to read additional details about those accuracy and error measures in "Additional Notes" (Sections 5.1 and 5.2).

TABLE 3. SPECIES CLASSIFICATION PERFORMANCE MATRIX.

		Classification (from the lidar inventory)										Total	PR
	Sp	Aw	Bw	Dp	Fb	Lt	Pb	Pl	Sb	Sg	Sw	**(row)**	
Reference /ground	Aw	48	3				3		1			55	87%
	Bw	2	15						1			18	83%
	Dp			0								0	0
	Fb				6	2						8	75%
	Lt	1				9		1	1		2	14	64%
	Pb	2	1				7					10	70%
	Pl	2				1		6				9	67%
	Sb				1	3		1	18		5	28	64%*
	Sg			2						2		4	50%
	Sw	2			7	4		1	4		45	63	71%*
Total (column)		57	19	2	14	19	10	9	25	2	52	209	
PC		84%	79%	0	43%*	47%*	70%	67%	72%	100%	87%		$\mathrm{P_o}$=75%*
PAve		86%	81%	0	55%*	55%*	70%	67%	68%*	67%	78%		(*z*=-1.94)
Sp distributions		χ^2 = 6.35 (*p*-value = 0.70)											

Note: species (sp) are defined in Table 1, "*" indicates significantly lower than 80% (the example accuracy threshold) at α = 0.05 (one-tailed), PR is the correct proportion relative to the reference (in row), PC is the correct proportion relative to the classification by inventory (in column), PAve is the pooled average of PR and PC, $\mathrm{P_o}$ is the overall accuracy for all species combined, and "Sp distributions" denotes the species frequency distributions (to be discussed later). The *z* value for $\mathrm{P_o}$ is the one sample proportion Z-test statistic against 80%. The χ^2 and the *p*-value (to be discussed later) evaluate the equivalence of the species frequency distributions from the reference (ground) and classification.

The notations for the four measures described above may not be immediately clear to some readers, but are necessary for generalization. To clarify the concepts further and demonstrate their calculations, a real world example obtained from a 2018-2020 airborne lidar inventory in the Forest Management Agreement (FMA) area of Canadian Forest Products Ltd. (Grande Prairie) is used (Forsite Consultants Ltd. 2020). Table 3 lists the frequency counts for the 10 species observed on the ground and identified by the inventory. Since such a table specifically addresses the categorical variable "species", it is also termed the "species classification performance matrix" – a special case of an integrated error matrix for species.

Based on the counts listed in Table 3, the overall species classification accuracy for all species combined is (from [2.1]):

$P_o = \sum_{i=1}^{k} n_{ii}$ /N = (48+15+0+6+9+7+6+18+2+45)/209 = 156/209 = 75%.

At the category (individual species) level, for instance, 48 aspens are correctly classified among a total of 55 ground observed aspens (references, row total). Therefore, the correct proportion relative to the reference for Aw is $PR_{aw} = n_{ii}/nr_i$ = 48/55 = 87% (from [2.2]). Similarly, for Sw, the correct proportion relative to the reference is PR_{sw} = 45/63 = 71% (45 white spruce trees are correctly classified among a total of 63 ground observed white spruce trees).

Since the total number of counts classified as Aw by the classification is 57 (column total), the correct proportion relative to the classification for Aw is $PC_{aw} = n_{ii}/nc_i$ = 48/57 = 84% (from [2.3]). Similarly, for Sw, the correct proportion relative to the classification is PC_{sw} = 45/52 = 87% (the total number of counts classified as Sw by the classification is 52).

The pooled average of the correct proportions for Aw is $PAve_{aw} = 2n_{ii}/(nr_i+nc_i)$ = (48+48)/(55+57) = 86%, which is obtained from [2.4]. Similarly, for Sw, the pooled average is $PAve_{sw}$ = (45+45)/(63+52) = 78%.

The PR and PC are two species-specific accuracy proportions (in percentages) calculated in two different ways, one relative to the reference in row (from the ground observations), and the other relative to the classification in column (from the inventory technique). As such, they can give very different accuracy assessments for a species. This needs to be clearly understood.

The PR represents the correct proportion of a species in relation to the total number of reference counts (in this case, ground counts) for that species. It indicates the proportion or probability that a ground species is correctly classified by the inventory, e.g., what percent of the white spruce trees in the ground sample were correctly classified as white spruce in the inventory?

The PC represents the correct proportion of a species in relation to the total number of counts classified as that species by the inventory. It indicates the proportion or probability that the species classified by the inventory actually represents the true species on the ground, e.g., what percent of the trees classified as white spruce in the inventory are actual white spruce trees on the ground?

The difference between PR and PC is caused by the misclassifications among the species. More specifically, it is caused by misclassifying the species of interest into other species (omission from the species of interest) and other species into the species of interest (commission error). It is important to recognize that either one of PR or PC only assesses one aspect of species-specific classification. It can be very misleading if only one of them is considered. Both must be looked at to avoid misinterpreting the accuracy information when assessing the accuracy of individual species classification from an inventory.

In practice, since either one of PR or PC may be misinterpreted or misunderstood, it can be very useful to have a single overall accuracy measure for each individual species present in an inventory, similar in concept to the overall accuracy measure for all species combined, but only for each individual species. There are varying ways to integrate the PR and PC values, or to derive other composite quantities and measures from the error matrix, through "different averaging methods", "discrete multivariate analysis techniques" that "balance" or "normalize" the original values, analysis of variance and different kappa statistics. We provide more details on many of them in "Additional Notes" (Section 5.2). The simplest and most obvious and effective approach for obtaining a single overall accuracy for each species is to just take the pooled or weighted average of the PR and PC values, as in PAve, which was used by Helldén (1980) and sometimes referred to as "the mean accuracy index" or "Helldén's mean accuracy index" (Rosenfield and Fitzpatrick-Lins 1986, Türk 2002, Liu et al. 2007, Stehman and Foody 2019). The pooled average expressed in PAve provides a unique and clearly understandable overall accuracy measure for each individual species in an inventory.

In many photogrammetric and remote sensing studies, an abridged table similar to Table 2 or Table 3 has often been called "confusion matrix" or "confusion table" (Story and Congalton 1986; Lillesand et al. 2015; Foody 2002, 2020), even though one of the earliest examples of such a table is called "error matrix" (Aronoff 1982). The concept of "confusion matrix" was introduced into remote sensing studies to quantify the confusion between categories (i.e., classes, species), and not the confusion it causes in the person trying to understand the matrix. It is very common to use a "confusion matrix" to represent the classification accuracy of remotely sensed data and maps, or use it as the basis for further analysis. There may not be

anything fundamentally wrong in calling such a table “confusion matrix”, except that it can be very confusing to most practitioners.

The confusions can be exacerbated in several ways: (1) when the confusion matrix uses the so-called "producer’s accuracy” to measure the “errors of exclusion (omission errors)”, and the "user’s accuracy” to measure the “reliability” or the “errors of inclusion (commission errors)”; (2) when the "producer’s accuracy” is labelled as (and confused with) the "user’s accuracy”, and the "user’s accuracy” is labelled as the "producer’s accuracy”; (3) when the "producer’s accuracy” is mixed/confused with “producer's risk” (the probability of incorrectly rejecting an acceptable map), and the “user’s accuracy” is mixed/confused with “consumer's risk” (the probability of accepting an inaccurate map); and (4) when the original values in the confusion matrix are "normalized" through "iterative proportional fitting”, which forces each row and column in the matrix to sum to one through iterations and changes to the original values in the rows and columns of the matrix. We will discuss these later for interested readers (in Sections 5.1 and 5.2).

We suggest the clear, intuitive terminology in Table 2 or Table 3 is used. We prefer to present and analyze the original data as they are, so that the results can be interpreted directly on the original rather than normalized, iteratively re-weighted or transformed data, which could distort the data and interpretation. We also purposely avoid the terms and idioms that may confuse many practitioners and some researchers, and think that “error matrix” or “classification performance matrix” (for any suite of categorical variables) is clearer, more pertinent and intuitive than the traditional “confusion matrix”. Interested readers who wish to find more details about why we choose to do so can read “Additional Notes” (Sections 5.1 and 5.2).

An added advantage of the classification performance matrix presented in Table 3 is that, it can be much more telling than the traditional confusion matrix. The classification performance matrix can not only provide the accuracy information for all species combined and by individual species, but also answer several questions related to the inventory, such as:

- Does the overall species classification accuracy meet the specified target (say, 80%, 85%, 90% or any other user specified percentages), when all species are combined?
- Do the accuracies of individual species meet the specified target?
- Do the accuracies for the same species (especially the leading species) differ significantly between the correct proportion relative to the reference (PR) and the correct proportion relative to the classification (PC)?
- Are the species frequency distributions from the reference and classification the same? In other words, can the two data sets, one obtained on the ground and the other from the inventory classification, be considered equivalent in representing the same species composition (i.e., species frequency distribution) for the same population (area of study)?

In practice, direct answers to the first two questions can be inferred directly from the P_o, PR, PC and PAve numbers listed in Table 3. For instance, since the overall accuracy P_o = 75%, we could say that it meets the accuracy target of 75% when all species are combined, but not 80%. For Aw, we could say that it meets the accuracy target of 80% because both PR and PC for Aw exceed 80% and PAve (Aw) = 86%.

However, for Sw, since PR_{sw} = 71%, PC_{sw} = 87% and PAve (Sw) = 78%, the answers are less certain, although we could say that the overall accuracy for Sw, PAve (Sw) = 78%, does not meet 80% but the accuracy relative to the classification, PC_{sw} = 87%, exceeds 80%. Where PR and PC are quite different, it is a signal that the species is being over- or under-called. In this case, Sw is being under-called so the PR is low. In the example table (Table 3), Fb is being over-called so the PC is low.

If statistically-based answers to all of the above questions are needed, we can implement statistical tests, provided that they are appropriately executed and interpreted. The answers can also be integrated into Table 3. Essentially the statistical tests allow us to compare the obtained and targeted accuracies more objectively, since it is not always possible to tell just by looking at them whether they are the same or different enough to be considered statistically significant. Statistical significance in this case means that the differences between the accuracies are not due to chance alone, but instead, they may be indicative of other factors at work.

Some readers may have noticed that Table 3 includes some additional numbers and symbols that have not been considered so far (e.g., the *z*-statistic, the chi-square (χ^2) statistic and associated *p*-value from statistical tests, and the star ‘*’ to indicate statistical significance). We will discuss them one-by-one below for interested readers.

Error Matrices for Species Groups and User-Defined Strata

Before moving on to discuss the additional numbers and symbols in Table 3 that can answer the above questions statistically, we want to mention that, Table 3 is explicit to individual species. Sometimes, for practical reasons, the accuracy of species classification from an inventory may only need to be assessed by the broad cover types (coniferous and deciduous) or by the species groups defined in Table 1. In those situations, for instance, Table 3 can be summarized into Table 4 if one is only interested in the accuracy for the broad cover types of coniferous and deciduous (for now assuming that the snags in Table 3 are coniferous – two of which are classified as deciduous and the other two (Dp) are coniferous).

TABLE 4. SPECIES CLASSIFICATION PERFORMANCE MATRIX FOR CONIFEROUS AND DECIDUOUS.

		Classification		Total (row)	Correct proportion re reference (PR)
	Species	**Coniferous**	**Deciduous**		
Reference/ground	Coniferous	119	7	126	94%
	Deciduous	2	81	83	98%
Total (column)		121	88	209	
Correct proportion re classification (PC)		98%	92%		P_o=200/209=96% (z=5.67)
Average of correct proportions by category (PAve)		96%	95%		
Species distributions		χ^2 = 0.25 (p–value = 0.62)			

Note: species (sp) are defined in Table 1, P_o is the overall accuracy, and "species distributions" denotes the species frequency distributions. The z value for P_o is the one sample proportion Z-test statistic against 80% (the example accuracy threshold). The χ^2 and the p-value (to be described later) evaluate the equivalence of the species frequency distributions from the reference (ground) and classification. The original species-specific data are listed in Table 3.

All species-specific formulas and calculations described earlier apply to Table 4, except now that there are only "two species", two categories (coniferous and deciduous) or two classes.

Intrinsically, the species classification performance matrix can be considered a part of the broader error matrices for an inventory or an image/map classification. Besides species, many inventories involve other categorical variables. Some (such as the AVI) also conduct a mixed land cover and land use classification. Relevant classification performance matrices for categorical variables from any inventory techniques can be constructed following the logic embedded in Tables 2-4. Depending on the analysis and the variable(s) of interest, the numbers listed in the cells (rows and columns) of the matrices can be counts or frequency proportions. They can also be areas, pixels, clusters, point clouds, stands or polygons, etc.

2.2 Crown delineation or stem segmentation accuracy for tree-based approach

For tree-based approach (as opposed to area-based approach), prior to assessing the accuracy of species classification, it is important to look at the accuracy of crown delineation or stem segmentation. This is because the overall accuracy of species classification for an individual tree-based inventory is comprised of two separate accuracy components: 1) crown delineation or stem segmentation accuracy; and 2) species classification accuracy.

The species classification accuracy in the form of the error matrix discussed above is derived based on the matched sample pairs of known ground species versus inventory predicted species (see more details in Section 2.3). Crown delineation accuracy is different and separate from the species classification accuracy – they are literally two very distinct processes. Crown delineation occurs first, and can result in several different outcomes. It can produce three types of errors:

1. **Missing error** – crown is completely missed or undetected during crown delineation.

2. **Under-counting error** – multiple crowns are incorrectly delineated into fewer crowns or a single crown, perhaps because they are blocked, invisible, cluttered, or clumped together. Inherently, under-counting errors can be considered missing errors, as they are caused by missing the crowns that should have been delineated and counted.

3. **Over-counting error** – a crown is incorrectly delineated into multiple crowns, perhaps because it is large and/or has an irregular shape. Sometimes a crown that does not exist is delineated, resulting in a "phantom" or a "ghost" crown. Phantom or ghost crowns are extra crowns that can be considered over-counting errors, as they are caused by over-counting the crowns that should not have been delineated and counted.

Table 5 lists the ground data and the corresponding inventory (lidar) data from two sample plots used to illustrate the calculations of the crown delineation errors. For simplicity, only tree number (Tree), tree species (Sp) and tree height (HT) are listed in Table 5.

TABLE 5. TREE-BASED GROUND AND INVENTORY DATA FROM TWO SAMPLE PLOTS (FOR ILLUSTRATION).

Plot	Ground			Inventory (lidar)			Note	In/Out
	Tree	Sp	HT	Tree	Sp	HT		(correct/incorrect)
1	G1	Sw	25.60	L1	Sw	25.75		In (correct)
1	G2	Aw	23.95	L2	Aw	23.97		In (correct)
1	G3	Pb	18.01	L3	Aw	20.97	Pb identified as Aw	In (incorrect)
1	G4	Aw	18.80				G4 missing	Out
1	G5	Aw	19.25	L4	Aw	19.54		In (correct)
1	G6	Sw	14.08				G6 missing	Out
1	G7	Fb	19.45	L5	Fb	20.78		In (correct)
1	G8	Sw	18.96	L6	Sw	17.91		In (correct)
1	G9	Sw	19.86	L7	Sw	19.80		In (correct)
1	G10	Aw	19.65	L8	Aw	19.35		In (correct)
2	G1	Sw	11.87	L1	Fb	11.77	Sw identified as Fb	In (incorrect)
2	G2	Aw	23.42	L2	Aw	23.56		In (correct)
2	G3	Aw	22.18	L3	Aw	23.30		In (correct)
2	G4	Pl	16.66	L4	Sb	18.88	Pl identified as Sb	In (incorrect)
2	G5	Pl	19.38	L5	Pl	19.03		In (correct)
2	G6	Sw	8.66				G6 missing	Out
2	G7	Aw	20.60	L6	Aw	18.23		In (correct)
2	G8	Sw	12.38	L7	Sb	13.39	Sw identified as Sb	In (incorrect)
2	G9	Aw	21.71	L8	Aw	20.37	Two Aw delineated as one tree	In (correct)
2	G10	Aw	17.14				(or, either G9 or G10 missing)	Out
2	G11	Sw	16.22	L9	Sw	19.50		In (correct)
2	G12	Aw	24.23	L10	Aw	23.18	One Aw delineated as two trees	In (correct)
2				L11	Pb	24.05		Out
2				L12	Bw	14.16	Non-existent “ghost” tree	Out
2	G13	Aw	22.68	L13	Aw	22.30		In (correct)

Note: Tree, Sp and HT denote tree number, tree species and tree height (m), respectively. Tree species are defined in Table 1. Shaded trees involve delineation errors. The last column “In/Out (correct/incorrect)” indicates two elements: if the tree is included (In) or excluded (Out) in assessing the species classification accuracy; and if the tree species is correctly (correct) or incorrectly (incorrect) identified by the inventory.

The “In/Out (correct/incorrect)” column in Table 5 indicates two elements: if the tree is included (In) or excluded (Out) in assessing the species classification accuracy (see later in Section 2.3); and if the tree species is correctly (correct) or incorrectly (incorrect) identified by the inventory. For plot 1 in Table 5, among the 10 ground trees, eight are delineated and two are missed by the inventory. Among the eight delineated trees, tree G3 (Pb) is misidentified as Aw, and all other tree species are correctly identified by the inventory. For plot 2 in Table 5, there are 13 ground trees. The inventory delineated 13 trees – but this does not mean that the delineation is perfect. The 13 delineated trees include:

- One missing tree (G6);
- Two trees (G9, G10) delineated as one tree (L8) – equivalent to under-counting one tree;
- One tree (G12) delineated as two trees (L10, L11) – equivalent to over-counting one tree;
- One non-existent tree delineated as a “ghost” tree (L12) – equivalent to over-counting one tree.

Apparently, for plot 2, the missing, under-counting and over-counting errors cancelled or balanced out when summed up, resulting in the delineated stems (crowns, trees) equaling to the total number of stems on the ground. This could be mistaken as a perfect delineation if a person only looks at the total number of ground versus delineated stems. There are other more convoluted crown delineation situations in which the crown delineation errors may cancel out. If not careful, they could lead to some “look good” but misleading inferences.

In general, when over-counting happens, there is an increased probability that the reference trees are classified correctly leading to a good PR value, but the many additional trees assigned to the class in error, result in a poor PC value. When under-counting happens, there is a reduced chance to get a good PR value because not enough trees of that species are called to match the trees in the ground sample. The PC value can still be good because the algorithm may be quite good at identifying clear examples of that species, so the few that do get called are mostly correct.

For tree-based approach, it is very important to at least take a look at the proportion of the stems that are missed by crown delineation.

However, in order to really understand the accuracy of crown delineation and avoid the crown delineation errors’ cancelling out problem, it is advisable that the three types of errors described above should be assessed separately. Table 6 lists the crown

delineation errors for the data in Table 5. Other more detailed statistics, such as the size class and species composition of the missed stems by the crown delineation, could also be assessed if needed for some specific studies (perhaps designed to search for causes, reasons, justifications and solutions).

TABLE 6. CROWN DELINEATION (OR STEM SEGMENTATION) ERRORS FOR THE DATA IN TABLE 5.

Plot	N_{ground}	N_{lidar}	Relative percent	Crown delineation error			Combined % error	Combined % correct
				Missing	Under-counting	Over-counting		
1	10	8	80.0%	2/10	0/10	0/10	2/10=20.0%	80.0%
2	13	13	100.0%	1/13	(2-1)/13	(3-1)/13	4/13=30.8%	69.2%
Sum	**23**	**21**	**91.3%**	**3/23**	**1/23**	**2/23**	**6/23=26.1%**	**73.9%**

Note: under-counting for plot 2 is resulted from two trees delineated as one tree, over-counting for plot 2 is resulted from one tree delineated as two trees and one non-existent tree delineated as a "ghost" tree. Definitions for other terms and variables are provided in the main text.

Where in Table 6: N_{ground} is the number of stems (crowns, trees) on the ground; N_{lidar} is the number of stems delineated by the inventory (lidar); Relative percent is N_{lidar} in relation to N_{ground} (Relative percent = N_{lidar}/N_{ground}); Missing, under-counting and over-counting are missing, under-counting and over-counting crown delineation errors, respectively (see Table 5); Combined % error is the summation of missing, under-counting and over-counting errors; Combined % correct is (100% – combined % error), which can be interpreted as the correct percentage of crown delineation for the combined samples.

The under-counting error for plot 2 is resulted from two trees (G9, G10) delineated as one tree (L8). The over-counting error for plot 2 is resulted from one tree (G12) delineated as two trees (L10, L11) and one non-existent tree delineated as a phantom or "ghost" tree (L12).

The relative percent in Table 6 is N_{lidar} in relation to N_{ground}. It is just a ratio between N_{lidar} and N_{ground}, not the correct % about crown delineation. For instance, there are cases where $N_{lidar} > N_{ground}$ (e.g., due to over-counting errors), which would result in a relative percent of greater than 100%. Had this relative percent been interpreted to be the correct % of crown delineation, it would mean that the correct % were greater than 100%. Obviously that does not make any sense. Only when there are no under-counting and over-counting errors, relative percent equals to the correct percentage of crown delineation (e.g., for plot 1, since under-counting error = 0 and over-counting error = 0, relative percent = combined % correct = 80.0%; But for plot 2, relative percent = 100%, while combined % correct = 69.2%, see Table 6).

Crown delineation accuracy is a critical consideration in determining the overall validity and usefulness of a tree-based approach. The errors in crown delineation should be appropriately quantified and assessed before the merits of a tree-based approach can be determined. Without conjointly considering crown delineation accuracy, the reported species classification accuracy can be inflated.

Often, however, crown delineation from canopy height models, canopy structures and point clouds can be difficult and challenging, even though considerable progresses and improvements have been reported in research (Yu et al. 2011, Wu et al. 2016, Mohan et al. 2017, Surový and Kuželka 2019, Coops et al. 2021, Prieur et al. 2022) and in operation (Strîmbu and Strîmbu 2015; Zhang et al. 2016, 2022; Yang and Kondoh 2020; Forsite Consultants Ltd. 2020). Crown delineation errors are not readily identifiable nor easily quantifiable from an automatic delineation process. To be able to identify and parse the errors into missing, under-counting and over-counting errors, some separate, more intelligent crown matching algorithms, or some visual matching and checking through an onerous manual process may be necessary. In any case, before implementing a tree-based approach as an inventory technique in operations, one needs to assess and understand the crown delineation or stem segmentation errors (at least the missing errors). Furthermore, without assessing and understanding the crown delineation errors, it is inadvisable to implement a tree-based approach as a data collection tool in place of measuring PSPs and TSPs on the ground. Besides their numerous other functions, ground-measured PSPs and TSPs are generally considered "the truth" and used as the reference standards for determining the accuracy and judging the validity of any forest inventory techniques.

While understanding the crown delineation accuracy is critical for tree-based approach, it is typically irrelevant for area-based approach. Area-based crown information (such as crown area, crown cover percent) can be obtained from an area-based approach. There is no crown delineation involved for individual trees in an area-based approach. Therefore, the crown delineation errors discussed above do not apply to area-based approach.

2.3 Further notes on species classification accuracy

The error matrix for species classification described in Section 2.1 is constructed based on the matched-pairs of known versus predicted tree species. To construct the error matrix, a set of samples (say, size *n*) is taken from a population. These samples with known ground species are compared to their respective predicted species from the inventory (in this case, from lidar). This means that the delineated individual crowns (stems, trees) by the inventory must first be matched correctly to the stems identified in the field, then compared and fed into the error matrix.

As an example to illustrate further, in order to calculate the species classification accuracy for the data in Table 5, the delineated individual stems by the lidar inventory must first be matched correctly to the stems measured on the ground. It can be seen from Table 5 that eight out of eight delineated trees for plot 1 can be matched with the ground trees, and 11 out of 13 delineated trees for plot 2 can be matched with the ground trees. The matched-pairs (a total of 19 from two plots) with known ground species and inventory predicted species are used to calculate the species classification accuracy. Results are shown in Table 7, where N_{ground} and N_{lidar} are the stem counts on the ground and from the inventory, respectively.

TABLE 7. SPECIES CLASSIFICATION ACCURACY FOR THE DATA IN TABLE 5.

Plot	N_{ground}	N_{lidar}	Matched-pairs	% of matched-pairs	Correct species prediction pairs	% Correct
1	10	8	8	8/8=100.0%	7	7/8=87.5%
2	13	13	11	11/13=84.6%	8	8/11=72.7%
Sum	**23**	**21**	**19**	**19/21=90.5%**	**15**	**15/19=78.9%**

Other variables appeared in Table 7 are defined as follows:

- Matched-pairs = number of matched ground-inventory pairs;
- % of matched-pairs = matched-pairs/N_{lidar}. It is the percentage of matched-pairs relative to N_{lidar};
- Correct species prediction pairs = number of correct species prediction pairs among matched-pairs;
- % Correct = correct species prediction pairs/matched-pairs.

As discussed before, the species classification accuracy is more commonly expressed in terms of the error matrix described in Section 2.1. For the example data in Table 5, the error matrix that corresponds to Table 7 is shown in Table 8.

TABLE 8. ERROR MATRIX FOR SPECIES CLASSIFICATION FOR THE DATA IN TABLE 5.

		Inventory (lidar)						Total	PR
	Species	Sw	Sb	Aw	Pb	Fb	Pl	**(row)**	
Ground (reference)	Sw	4	1	0	0	1	0	6	67%
	Sb	0	0	0	0	0	0	0	0
	Aw	0	0	9	0	0	0	9	100%
	Pb	0	0	1	0	0	0	1	0
	Fb	0	0	0	0	1	0	1	100%
	Pl	0	1	0	0	0	1	2	50%
Total (column)		4	2	10	0	2	1	19	
PC		100%	0	90%	0	50%	100%		
Pave		80%	0	95%	0	67%	67%		P_o=78.9%

Note: species are defined in Table 1, PR is the correct proportion relative to the ground (reference), PC is the correct proportion relative to the classification (prediction) by the inventory, PAve is the pooled average of PR and PC, and P_o is the overall accuracy for all species combined.

The overall accuracy listed in Table 8 for all species combined (P_o=78.9%) is identical to the % Correct in Table 7 when all data are combined. The error matrix (Table 8) is preferred over Table 7 because it also provides more detailed error/accuracy statistics for individual species. The foundation for Table 7 and Table 8, however, is the same. They both must be derived based on only the matched-pairs of known ground versus inventory predicted species. Unmatched pairs and any associated missing errors and species misidentification errors in a tree-based approach can be assessed during the crown delineation process (Section 2.2), but they are not accounted for by any means in the error matrix for species classification, which requires matched-pairs for the calculation.

In the “purest theoretical sense”, the data assessed for the species classification accuracy should be free of the delineation errors to avoid the inflation or deflection of the accuracy statistics. Therefore, ideally the shaded trees in Table 5, which involve delineation errors, may not be used in assessing the species classification accuracy. In practice, however, so long as there is

a match-able pair resulted from the delineation, the matched-pair should be used in the evaluation of species classification accuracy.

For example, for plot 2 in Table 5, two Aw trees (G9, G10) are incorrectly delineated as one Aw tree (L8). The delineation involves an under-counting error. However, since the delineated Aw can be matched with one of the two Aw trees, it should not be excluded or thrown away in assessing the species classification accuracy (throw away the Aw in this case would underestimate the accuracy). For the same logic, if the two Aw trees are delineated as one Pb tree, it should not be thrown away either, as it can still be paired with one of the two Aw trees, notwithstanding that the delineated species is incorrect (throw away the Pb in this case would overestimate the accuracy).

In a different scenario (see Table 5), assume that one Aw tree in plot 2 (G12) is delineated as three trees: one Aw (L10), one Pb (L11) and one Bw (L12). The delineation involves some over-counting errors. Since at least one of the three delineated trees can be paired with the Aw tree (G12), they should not be thrown away in assessing the species classification accuracy.

A more intricate practical problem arises when there are over-counting errors. In the above example, one Aw is delineated as three trees of different species: one Aw, one Pb and one Bw. Which one of the three delineated trees should be matched with the ground reference? The varied choices will not only impact the assessment of the species classification accuracy, but also other accuracies related to other species-specific variables (e.g., tree height, stand height and stand density).

Readers and agencies can develop their own “best practices” or hierarchies for the matches if over-counting occurs. These “best practices” are beyond the scope of this study. They are dependent on the consideration of which variable is the most important for a specific study. As an example, one could implement the following order of precedence for consistent pair matching:

1. The same species regardless of the other variables;
2. The closest heights of the same species;
3. The closest heights regardless of the species;
4. The closest horizontal distance between the ground and delineated stems.

For instance, following this order of precedence, for the ground Aw delineated as Aw, Pb and Bw by the inventory, the ground Aw is matched with the delineated Aw regardless of the other variables.

If the ground Aw is delineated as Aw, Aw and Bw by the inventory, the delineated Aw whose height is closer to that of the ground Aw is matched with the ground Aw.

If the ground Aw is delineated as Pb, Pb and Bw by the inventory, the delineated tree whose height is closest to that of the ground Aw is matched with the ground Aw.

If the ground Aw is delineated as Pb, Pb and Bw by the inventory, and the spatial locations of all four trees are known, the delineated tree that is the closest (in terms of the horizontal distance) to the ground Aw is matched with the ground Aw.

Regardless of the developed “best practices” for the matches, fundamentally, the error matrix for species classification must be derived based on the matched-pairs of known versus predicted species. The overall accuracy of species classification for any tree-based approach must be evaluated simultaneously based on the error matrix for species classification and the crown delineation accuracy. Without knowing the crown delineation or stem segmentation accuracy, the reported high species classification accuracy may not mean much in judging the quality of an inventory technique.

2.4 One sample proportion z-test

For practical and regulatory purposes, an accuracy target or an acceptable accuracy threshold is frequently established to determine, for instance, whether the overall species classification accuracy meets the threshold of 80% (or any other “reasonable” numbers one may select, such as 70%, 75%, 85%, 90%, 95%, etc.). While the concept of establishing such a threshold has its obvious merit, in broad forest inventory applications it can be more difficult to achieve agreement on what constitutes a reasonable, a good, an average, or a poor classification accuracy. In some cases 80% may represent a reasonable or a good accuracy whereas in other cases it may not. The problem can become more complex owing to, among many other quantitative and non-quantitative factors, different goals and objectives, types of variables and data, the variable costs of making class-specific errors (Gergel et al. 2007), and the sample sizes involved in an inventory.

Here, we will not make a judgment on what may constitute the "best" threshold (we will leave it for others to decide), but rather, simply use 80% as an example threshold to demonstrate the methodology, which conversely, can assist the establishment of a "realistic and achievable" accuracy threshold for a classification in practice. Readers could choose other suitable subjective thresholds where appropriate and justifiable. The methodology remains the same.

Once the accuracy threshold of 80% is specified, the one sample proportion Z-test (or simply, one sample Z-test) can be applied to answer the two questions related to an inventory technique: 1) if the overall species classification accuracy meets the specified threshold; and 2) if the accuracies of individual species classifications meet the specified threshold.

The one sample proportion Z-test compares an observed sample proportion to a threshold or a target proportion chosen by a user. It is implemented by computing the following Z-test statistic:

[2.7] $$z = \frac{p_1 - p_t}{\sqrt{\frac{p_t(1-p_t)}{n}}}$$

where z is the test statistic, p_1 is the accuracy proportion from the samples for the category (e.g., species), p_t is the specified threshold proportion, and n is the total number of samples.

The null (H_0) and alternative (H_a) hypotheses for the one sample proportion Z-test are:

H_0: the sample proportion (p_1) is equal to the specified threshold proportion ($p_1 = p_t$).

H_a: p_1 is not equal to the specified threshold proportion ($p_1 \neq p_t$) (two-tailed or two-sided).

H_a: p_1 is greater than the specified threshold proportion ($p_1 > p_t$) (one-tailed, right sided).

H_a: p_1 is smaller than the specified threshold proportion ($p_1 < p_t$) (one-tailed, left sided).

At the specified significance level of $\alpha = 0.05$ (used throughout this study), for the two-tailed test of $p_1 \neq p_t$, the critical value is ±1.96. The decision rule is that, if the calculated z is $\leq$ -1.96 or $\geq$ +1.96 (i.e., if $|z| \geq 1.96$), reject the null hypothesis. For the one-tailed test of $p_1 > p_t$, the critical value is +1.645. The decision rule is that, if $z \geq 1.645$, reject the null hypothesis. For the one-tailed test of $p_1 < p_t$, the critical value is -1.645. The decision rule is that, if $z \leq -1.645$, reject the null hypothesis.

As an example, based on the number of trees data in the species classification performance matrix (Table 3), the overall species classification accuracy from the inventory is, $p_1 = 156/209 = 0.7464$, where 156 are the total number of correct predictions and 209 are the total number of samples. Hence,

[2.8] $$z = \frac{0.7464 - 0.80}{\sqrt{\frac{0.80(1-0.80)}{209}}} = -1.9368 = -1.94.$$

For the two-tailed test, since the calculated z statistic is between -1.96 and 1.96, it means that we fail to reject the null hypothesis of $p_1 = p_t = 0.80$ (at $\alpha = 0.05$). That is equal to say that, statistically, the overall accuracy percent obtained from the inventory (75%) is not different from the 80% threshold. In other words, even though we only achieved 75% accuracy from the samples, we can claim that statistically it meets the specified accuracy threshold of 80% when all species are combined.

For the one-tailed test of $p_1 < p_t$, since the calculated $z = -1.94 < -1.645$, it means that the null hypothesis of $p_1 = p_t = 0.80$ is rejected. That is equal to say that, statistically, the overall accuracy percent obtained from the inventory (75%) is lower than the 80% threshold. In other words, we cannot claim that the obtained accuracy (75%) meets the specified accuracy threshold of 80% when all species are combined.

The conflicting inferences obtained above from the two-tailed test and one-tailed test are not uncommon when a test statistic is near the critical borderlines. They mean that additional samples and/or studies are needed to achieve a more conclusive result. Hypothesis testing is highly dependent on the sample size (and the way the samples are collected). The smaller the sample size the less likely one will find a difference. Sample size and the way the samples are collected (i.e., sampling method) are very important considerations in developing and assessing forest inventory techniques. For our purpose, since we are mostly interested in if the sample proportion is equal to, greater than, or smaller than the specified threshold proportion, the one-tailed tests are more appropriate.

The *one sample* proportion *Z-test* is also conducted for each species in Table 3 (recognizing that the total number of samples for each species is different for "reference" and "classification"). For instance, the PR (reference) for Aw with an $n = 55$ has a $z = (48/55 - 0.80)/\sqrt{0.80(1-0.80)/55} = 1.35$. The PC (classification) for Aw with an $n = 57$ has a $z = (48/57 -$

0.80)/$\sqrt{0.80(1-0.80)/57}$ = 0.79. Both are within [-1.645, 1.645]. Therefore, both can be considered (statistically) equivalent to 80%.

The pooled average of the correct proportions for Aw from [2.4] is $PAve_{aw}$ = 86%. This pooled average for Aw corresponds to a *z* = ((48 + 48)/(55 + 57) – 0.80)/$\sqrt{0.80(1-0.80)/(55+57)}$ = 1.51, which is within [-1.645, 1.645]. Therefore, it is also (statistically) equivalent to 80%.

Table 9 lists the calculated *z* values for Aw and for all other species based on the results in Table 3 (except for Dp where PR=PC=PAve=0). The *z* values for Sb and Sw relative to the reference (PR), the *z* value for Fb and Lt relative to the classification (PC), and the *z* values for pooled averages for Fb, Lt and Sb, are outside the critical boundaries of [-1.645, 1.645] for the one-tailed test. Therefore, for these species, the accuracy percentages listed in PR, PC and PAve in Table 9 did not meet the specified threshold of 80% for the one-tailed test. Interested readers can make appropriate inferences for the two-tailed test based on the z values listed in Table 9 (the only difference is that the PR for Sw is insignificant for the two-tailed test).

TABLE 9. ONE SAMPLE PROPORTION Z-TEST FOR INDIVIDUAL SPECIES IN TABLE 3.

Accuracy	Aw	Bw	Fb	Lt	Pb	Pl	Sb	Sg	Sw
PR	87%	83%	75%	64%	70%	67%	64%*	50%	71%*
	(1.35)	(0.35)	(-0.35)	(-1.47)	(-0.79)	(-1.00)	(-2.08)	(-1.50)	(-1.70)
PC	84%	79%	43%*	47%*	70%	67%	72%	100%	87%
	(0.79)	(-0.11)	(-3.47)	(-3.56)	(-0.79)	(-1.00)	(-1.00)	(0.71)	(1.18)
PAve	86%	81%	55%*	55%*	70%	67%	68%*	67%	78%
	(1.51)	(0.16)	(-2.98)	(-3.66)	(-1.12)	(-1.41)	(-2.20)	(-0.82)	(-0.47)

Note: species are defined in Table 1, PR is the correct proportion relative to the ground reference, PC is the correct proportion relative to the inventory classification, and PAve is the pooled average of the PR and PC for each species. Actual values of PR, PC and PAve are listed in Table 3. The values in parentheses are *z* values from the one sample proportion Z-test against a threshold of 80%. The values marked with a "*" indicate statistical significance (at α = 0.05) against the specified threshold of 80% for the one-tailed test.

The one sample proportion Z-test results obtained for all species combined (*z* = -1.94 from [2.8]) and for individual species (Table 9) can be incorporated into the species classification performance matrix (Table 3). Other than the percentage values marked with a significance sign "*" (at α = 0.05, one-tailed) in Table 3, all other values listed in PR, PC and PAve are (statistically) equivalent to or larger than the specified threshold of 80% (except for Dp where PR=PC=PAve=0). Therefore, Table 3 in fact answered the questions of whether the overall and individual species classification accuracies met the specified threshold of 80% (of course, as a note of caution, we cannot rely too much on any statistics for any species with a sample size of 10 or less. The "best" sample size is dependent on many factors, chief among them, the cost and the time, the variations and the number of variables involved, the required precision and the allowable errors (or the acceptable risks), plus the objective, range and scope of a study).

Several additional observations can be made from the one sample proportion Z-test results in Table 9 (which were incorporated into Table 3), all at α = 0.05, one-tailed:

1). If the interpretation is based on the correct proportion relative to the ground reference (PR) without conducting the one sample Z-test, only two species, Aw and Bw, meet the 80% threshold. Based on the one sample Z-test, all species except Sb and Sw meet the 80% threshold.

2). If the interpretation is based on the correct proportion relative to the inventory classification (PC) without conducting the one sample Z-test, only three species, Aw, Sg and Sw, meet the 80% threshold. Based on the one sample Z-test, Fb and Lt do not meet the 80% threshold, while all other species including Sb meet the 80% threshold.

3). Note that using PR or PC changed the conclusions for some species (e.g., for Sb, Sw, Fb and Lt). This is expected because PR and PC can produce two very different quantities due to the misclassifications among the species by the inventory. Either the PR or the PC only looks at one aspect of a classification for a species, although sometimes there are cases where it may be desirable to weigh focus on either one of the measures depending on the purpose of the map being inventoried.

4). The PAve calculated according to [2.4] or [2.5] combines both PR and PC. It provides a singular overall accuracy measure for each individual species. Together with the one sample proportion Z-test, it also answers the question of whether the overall accuracy for each individual species meets the specified accuracy threshold. Based on the PAve and the results in Table 9, it is very clear that Fb, Lt and Sb, with PAve values of 55%, 55% and 68%, respectively, did not meet the specified threshold of

80% whereas all other species met. The PAve can be more meaningful in practice if one is only interested in the overall accuracy for each individual species.

The one sample proportion Z-test is a generic test that can evaluate a sample proportion against any target proportion. It can also be used to evaluate species composition proportions in terms of number of trees or crown cover areas for individual species.

2.5 Two sample proportion z-test

When judging the accuracies of different inventory techniques on the classification of individual species, the following questions often arise in one way or the other:

- Do the accuracies for the same species (especially the leading species) differ from different inventory techniques?
- How can we determine if different inventory techniques produce the same or different accuracy proportions for the same species in interest?

For instance, assuming that the error matrix in Table 3 is from inventory technique 1 (T1). For T1 the correct proportion relative to the ground reference for Aw is PR_{aw1} = 87% (from a sample size of n_1 = 55). Assuming also that from a different inventory technique 2 (T2), the correct proportion relative to the ground reference for Aw is PR_{aw2} = 78% (from a sample size of n_2 = 208). Are these two accuracies for Aw from T1 and T2 the same or different?

In practice, judging from the absolute PR values, we may say that for Aw, T1 appears more accurate than T2 (because PR_{aw1} = 87% > PR_{aw2} = 78%).

If a statistically-based answer to the above question is needed, the two sample proportion Z-test (or simply, two sample Z-test) can be implemented. The two sample proportion Z-test evaluates whether the difference between two proportions from two techniques, two methods or two samples for a categorical variable is statistically significant at the specified α level. It calculates the following test statistic:

[2.9] $$z = \frac{p_1 - p_2}{\sqrt{p(1-p)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

where z is the test statistic, p_1 is the accuracy proportion from technique 1 (or sample 1), p_2 is the accuracy proportion from technique 2 (or sample 2), n_1 is the number of samples from technique 1, n_2 is the number of samples from technique 2, and p is the pooled average accuracy proportion calculated by:

[2.10] $$p = \frac{p_1 n_1 + p_2 n_2}{n_1 + n_2} = \frac{k_1\ (\text{correct count from method 1}) + k_2\ (\text{correct count from method 2})}{n_1 + n_2}$$

The null and alternative hypotheses for the two sample proportion Z-test are:

H_0: the accuracy proportions from two techniques (T1 and T2) are the same ($p_1 = p_2$).

H_a: the accuracy proportions from two techniques are different ($p_1 \neq p_2$) (two-tailed).

H_a: the accuracy proportion from T1 is greater than that from T2 ($p_1 > p_2$) (one-tailed, right sided).

H_a: the accuracy proportion from T1 is smaller than that from T2 ($p_1 < p_2$) (one-tailed, left sided).

Since the two sample proportion Z-test statistic is a z-score, at the specified significance level of α = 0.05, the following decision rules apply:

- For the two-tailed test of $p_1 \neq p_2$, if the calculated z value from [2.9] is $\leq$ -1.96 or $\geq$ +1.96 (i.e., if $|z| \geq 1.96$), the null hypothesis is rejected, which means that there is a significant difference between the two accuracy proportions for the same species from two techniques. Otherwise, there is no significant difference between the two accuracy proportions.
- For the one-tailed test of $p_1 > p_2$, the critical value is +1.645. Hence, if the calculated $z \geq 1.645$, the null hypothesis is rejected, which means that the accuracy proportion from technique 1 is greater than that from technique 2.
- For the one-tailed test of $p_1 < p_2$, the critical value is -1.645. Hence, if the calculated $z \leq$ -1.645, the null hypothesis is rejected, which means that the accuracy proportion from technique 1 is smaller than that from technique 2.

For the above example for Aw from two inventory techniques (T1 and T2), p_1 = 87%, n_1 = 55, p_2 = 78% and n_2 = 208. Therefore, p = 0.7988 (from [2.10]) and z = 1.481 (from [2.9].

The calculated z value is between -1.645 and +1.645, which means that the null hypothesis of $p_1 = p_2$ is not rejected regardless of whether the one-tailed or two-tailed test is conducted. It suggests that, statistically, the accuracy proportions for Aw, PR_{aw1} = 87% and PR_{aw2} = 78% from two inventory techniques, can be considered equivalent to each other in a repeated sampling sense. The apparent difference between PR_{aw1} and PR_{aw2} is within the acceptable sample variation.

Following the same logic, the two sample proportion Z-test can be applied to any other species from different inventory techniques to determine if the inventory techniques produce the same or different accuracy proportions for the species in interest. For instance, it can be used to assess whether the leading species call from AVI polygon is the same as the one created by aggregated individual tree inventory data. In fact, the two sample proportion Z-test can be used to compare any two proportions obtained from two different techniques, methods, areas or user-defined strata.

3 Assessing species composition or species frequency distribution

The idea imbedded in the formulation of the species classification performance matrix (Table 3) focused on the species classification accuracies in proportions or percentages from an inventory technique. They apply to individual species and all species combined. However, they did not consider the overall accuracy in terms of the frequency distributions of the species as a whole (i.e., species compositions) from the ground and inventory. Therefore, they did not answer the question of whether the species frequency distributions (species compositions) from the ground and inventory are the same or different.

For instance, for the data in Table 3, the species frequency distributions from “correct”, “ground” and “inventory” are shown in Figure 1, where “correct” denotes the counts of correctly identified species by the inventory (the shaded diagonal values in Table 3), “ground” denotes the total species counts from the ground reference (the row totals in Table 3), and “inventory” denotes the total species counts identified by the inventory (the column totals in Table 3).

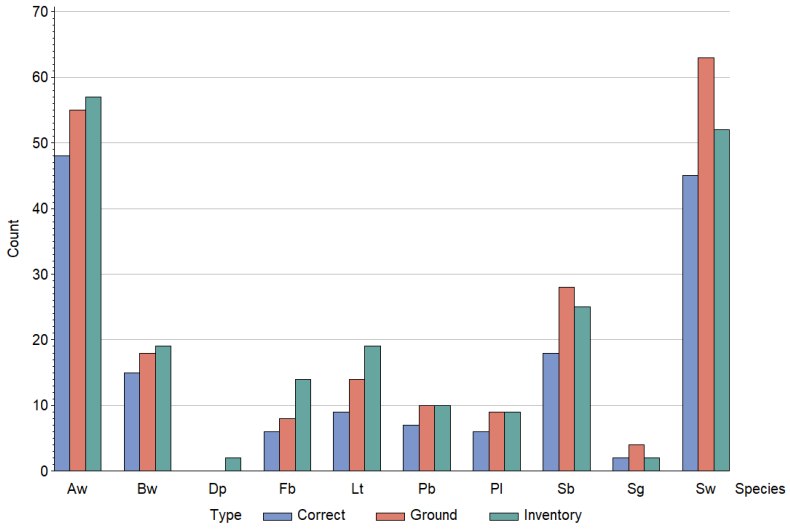

Figure 1. Species frequency distributions for the data in Table 3, where “correct” denotes the counts of correctly identified species by the inventory, “ground” denotes the total species counts from the ground reference, and “inventory” denotes the total species counts identified by the inventory.

Are the species frequency distributions between the ground and inventory the same? What about the species frequency distributions between the correct and the ground? How can we compare the ground measured species composition or a reference species composition to a species composition derived from a lidar/satellite inventory? How can we compare the species compositions from different inventory techniques? These can be critically important considerations in determining the validity of an inventory and comparing different inventory techniques.

Since the answer to the question of whether the species frequency distributions are the same involves the categorical variable “species”, the chi-square test and Fisher's exact test are appropriate. Both tests apply to categorical variables (also referred to as discrete, qualitative, discontinuous, nominal, or ordinal variables if there is a clear ordering of the categories).

Before moving onto discussing the chi-square test and Fisher's exact test, some brief clarifications on the terminologies related to “frequencies”, “frequency numbers”, “frequency proportions” and “frequency distributions” are helpful, as there appear to be some ambiguities and confusions about these terms in the literature. Most importantly, as will be shown later, the chi-square test and Fisher's exact test (and the Kolmogorov-Smirnov test for continuous variables, to be discussed later), only apply to the specifically defined frequency distributions or frequency proportions.

3.1 Clarification on frequency related terminologies

Often, the word “frequency” can have two different interpretations:

- Frequency may mean “frequency count” in actual number (or “frequency number”);

- Frequency may mean “frequency proportion” (frequency proportions sum up to 1 or 100%).

Both interpretations (and the word itself) are closely related to “cumulative frequency”, which can mean “cumulative count” or “cumulative proportion”. Table 10 uses an example to explicitly define and illustrate the differences among the terminologies related to different types of frequencies, where species 1-5 represent different species.

TABLE 10. DEFINITIONS AND EXAMPLES OF TERMINOLOGIES RELATED TO FREQUENCY AND FREQUENCY DISTRIBUTION.

Species	Frequency count	Frequency proportion	Cumulative count	Cumulative proportion
Species 1	3	0.3 (3/10)	3	0.3 (3/10)
Species 2	1	0.1 (1/10)	4	0.4 (4/10)
Species 3	4	0.4 (4/10)	8	0.8 (8/10)
Species 4	0	0 (0/10)	8	0.8 (8/10)
Species 5	2	0.2 (2/10)	10	1 (10/10)

Note: “frequency” may refer to “frequency count” or “frequency proportion”. In this study, “frequency distribution” refers to “frequency proportion” or “cumulative proportion”, but not “frequency count”, nor “cumulative count”.

The importance of the apparently simple and trivial differences among the four terms in Table 10 (frequency count, frequency proportion, cumulative count and cumulative proportion) will become clear later. It is beyond just the “semantics”. It is also beyond just simply choosing one to one’s preference and using it consistently. Moreover, in day-to-day usages, the four terms in Table 10 have also been loosely referred to as “frequency distributions”, or simply “distributions”. This has caused some confusions, as the results of the three tests to be discussed in this study (the chi-square test, Fisher's exact test and the Kolmogorov-Smirnov test), do not apply to all four terms, nor to the broadly and vaguely defined “distributions” although they have been frequently (and mistakenly) thought to apply.

In this study, the four terms are spelled out explicitly wherever needed. In addition, species frequency distributions refer specifically and exclusively to the species frequency proportions across different species (as in column 3, Table 10). These clarifications are necessary for the interpretations of the results from the three tests. We will provide more details on them later in Section 5.4.

3.2 The chi-square test for categorical variables

To evaluate how accurate an inventory technique is in representing the actual forest characteristics and conditions on the ground, we often collect a set of data from the ground and compare it to the classification from the inventory on the same landbase (i.e., the “paired” landbase). One of the most important questions related to such comparison is: are the species frequency distributions from the ground and inventory the same? In other words, can the two data sets from the ground and inventory be considered equivalent in representing the same species frequency distribution (i.e., species composition) for the same landbase?

To answer this question, either the chi-square test or Fisher's exact test can be used, depending on the “data conditions” to be discussed below. Methodologically both the chi-square test and Fisher's exact test can evaluate data sets of any dimensions, but here our focus is on the two dimensional (bivariate) data sets from the “ground” and “inventory”.

The Chi-Square Test

The chi-square (χ^2) test evaluates if two sets of data with unknown distributions have the same frequency distributions as each other, or if they come from the same frequency distribution with the same frequency proportion for the categorical variable involved. To illustrate the computations involved in the chi-square test, species frequency counts from Table 3 (or Figure 1) for “correct” and “ground” are listed in Table 11, where “correct” refers to the counts of correctly identified species by the inventory, and “ground” refers to the total species counts from the ground reference. Readers can also use the counts for “inventory” and “ground” from Table 3 for the computations (see below). The logic is the same.

To describe the computations in Table 11, the null and alternative hypotheses are specified first:

H_0: The species frequency distributions from the ground and correct inventory counts are the same.

H_a: The species frequency distributions from the ground and correct inventory counts are different.

The chi-square test is then conducted by calculating the following chi-square test statistic:

[3.1] $\chi^2 = \sum_i^r \sum_j^c \frac{(O_{ij}-E_{ij})^2}{E_{ij}}$

where r is the number of rows and c is the number of columns in the bivariate table, O_{ij} is the observed cell value, E_{ij} is the expected cell value ("expected" if the null hypothesis is true**),** i = 1, 2, ⋯, r and j = 1, 2, ⋯, c. The E_{ij} for each cell in the table is calculated by:

[3.2] $E_{ij} = \frac{Row\ i\ total \times Column\ j\ total}{Grand\ total\ (\text{row totals or column totals})} = \frac{n_{i\cdot} \times n_{\cdot j}}{N}$

where $n_{i\cdot}$ is the row i total, $n_{\cdot j}$ is the column j total and N is the grand total. The degrees of freedom (*df*) for the calculated χ^2 statistic is (r−1)×(c−1). If the calculated χ^2 statistic is greater than the critical value from the chi-square distribution, the null hypothesis is rejected.

TABLE 11. AN EXAMPLE OF THE CHI-SQUARE TEST FROM THE DATA IN TABLE 3.

Species	Ground (O_1)	Correct (O_2)	Row total	E_1	E_2	$(O_1 - E_1)^2/E_1$	$(O_2 - E_2)^2/E_2$
Aw	55	48	103	58.978	44.022	0.268	0.359
Bw	18	15	33	18.896	14.104	0.042	0.057
Dp	0	0	0	0.000	0.000	.	.
Fb	8	6	14	8.016	5.984	0.000	0.000
Lt	14	9	23	13.170	9.830	0.052	0.070
Pb	10	7	17	9.734	7.266	0.007	0.010
Pl	9	6	15	8.589	6.411	0.020	0.026
Sb	28	18	46	26.340	19.660	0.105	0.140
Sg	4	2	6	3.436	2.564	0.093	0.124
Sw	63	45	108	61.841	46.159	0.022	0.029
Total	**209**	**156**	**365**	**209**	**156**	**0.609**	**0.816**

Note: "ground" (total species counts from the ground reference) and "correct" (counts of correctly identified species by the inventory) are from Table 3. The calculations follow [3.1] and [3.2], where the number of species (rows) is i = 1, 2, ⋯, 9 (since Dp =0) and the number of columns is j = 1, 2, O_1 and O_2 denote observed values and E_1 and E_2 denote the corresponding expected values. See main text for step-by-step computations.

Step-by-step computations for the chi-square test are shown in Table 11. For instance:

$E_{11} = \frac{n_{i\cdot} \times n_{\cdot j}}{N} = \frac{103 \times 209}{365} = 58.978$ $\quad E_{12} = = \frac{103 \times 156}{365} = 44.022$ $\quad E_{92} = = \frac{108 \times 156}{365} = 46.159$

Notice that the sums of the observed values are equal to the sums of the expected values in each row and column of the table. The calculated chi-square test statistic from [3.1] is the summation of the numbers from the last two columns of Table 11, which is χ^2 = 1.425, with a *df* of (r−1)×(c−1) = (9-1)×(2-1) = 8 (Dp missing in both counts so there are nine species in the 9×2 table). This χ^2 = 1.425 corresponds to a *p*-value of 0.9939, which is greater than 0.05. Therefore, the null hypothesis is not rejected. This means that there is no significant difference between the frequency distribution of the species correctly identified by the inventory and that from the ground reference. In other words, the two species frequency distributions from the ground and correctly identified by the inventory are statistically the same. [Note: the *p*-value for the chi-square test is typically automatically outputted by any usable statistical software. It can also be "manually" calculated if one knows how to calculate the cumulative density function (CDF). A manual calculation program and an automatic program written in SAS are available to interested readers].

Similarly, the chi-square test can be conducted to evaluate if the species frequency distributions in Table 3 from the "ground" and "inventory" are the same. The computations follow those demonstrated in Table 11 (with the "correct" column replaced by the "inventory" (classification) counts from Table 3). The calculated χ^2 = 6.35, with a *df* of (r−1)×(c−1) = (10-1)×(2-1) = 9 (Dp has two observations in "inventory" so there are ten species in the 10×2 table). This χ^2 = 6.35 corresponds to a *p*-value of 0.7049, which is greater than 0.05. Therefore, the null hypothesis of equivalent species frequency distributions from the ground and inventory is not rejected. This means that the two species frequency distributions from the ground and inventory follow the same common frequency distribution and there is no significant difference between the two species frequency distributions. This is a good indication that as a whole, if we can ignore the stem segmentation errors (i.e., if we ignore the stems not counted/included in Table 3), the species frequency distribution from the inventory matches reasonably well with the species frequency distribution from the ground reference. This result can be interpreted from the χ^2 statistic and the accompanying *p*-value listed at the bottom of the species classification performance matrix (Table 3).

Condition for Using the Chi-Square Test

The chi-square test can be sensitive to small sample sizes because it is based on an approximation approach. For instance, for a 3×2 matrix illustrated in Table 12 with three species from the ground (G) and inventory (I), following the above example,

the chi-square test statistic is calculated to be χ^2 = 6.11, with a *p*-value = 0.0471 (<0.05), which indicates that the species frequency distributions from the ground and inventory are significantly different. However, since 83% (5 out of 6) of the cells in Table 12 have expected counts less than 5, the chi-square test may not be a valid test. An automatic warning message usually accompanies the output of the test in these situations by any usable statistical software.

TABLE 12. AN ILLUSTRATION OF A CHI-SQUARE TEST FOR THE DATA FROM GROUND AND INVENTORY.

Species	Ground (G)	Inventory (I)	Expected (G)	Expected (I)	Chi-square
Aw	4	2	6×12/20 = 3.60	6×8/20 = 2.40	χ^2 = 6.11
Bw	3	6	9×12/20 = 5.40	9×8/20 = 3.60	(*p*-value = 0.0471)
Sb	5	0	5×12/20 = 3.00	5×8/20 = 2.00	

For practical purposes, the rule of thumb is that if more than 20% of the cells in a table or matrix whose expected values are less than five, the chi-square test may not be a valid test (Cochran 1954). In a case like this, a correction to the chi-square test may be needed (such as Yates correction, which calculates the chi-square statistic as $\chi^2 = \Sigma\Sigma(|O_{ij} - E_{ij}| - 0.5)^2/E_{ij}$). But the best option is to implement Fisher's Exact Test.

3.3 Fisher's exact test for categorical variables

In the literature, Fisher's exact test has also been called "Freeman-Halton test", "Fisher-Freeman-Halton test" and "Fisher-Irwin exact test" (Armitage et al. 2002, Stokes et al. 2012, Agresti 2013, SAS Institute Inc. 2020). Since Fisher first came up with the idea and the exact method for the 2×2 matrix (also known as the 2×2 contingency table or simply 2×2 table), we still refer it simply as Fisher's exact test.

The details involved in Fisher's exact test can look very complex and confusing. Interested readers may wish to read two articles by Freeman and Halton (1951), who extended Fisher's method to general *r*×*c* matrices of any number of rows (*r*) and columns (*c*), and Mehta and Patel (1983), who developed the network algorithm to provide faster and more efficient computations for general *r*×*c* matrices. The notations alone in these and many articles on Fisher's exact test are, however, a bit of a deterrent for most non-statisticians in forestry. Here we will explain Fisher's exact test in simpler terms and use an example to illustrate the computations.

For an *r*×*c* matrix with any number of rows and columns, Fisher's exact test is implemented by first calculating the hypergeometric probability of the observed matrix, then the hypergeometric probabilities of all other possible matrices of nonnegative integers "conditional on the observed row and column totals" (i.e., with the row and column totals of all other possible matrices identical to those from the observed matrix), using:

[3.3] $$P_h = \frac{\prod_{i=1}^{r} R_i! \prod_{j=1}^{c} C_j!}{N! \prod_{ij} a_{ij}!} = \frac{(R_1!R_2!\cdots R_r!)\times(C_1!C_2!\cdots C_c!)}{(\sum_{i=1}^{r} R_i)!\times(a_{11}!a_{12}!\cdots a_{1c}!\times a_{21}!a_{22}!\cdots a_{2c}!\times\cdots\times a_{r1}!a_{r2}!\cdots a_{rc}!)}$$

where P_h is the so-called "hypergeometric probability" of a matrix with the given row and column totals, R_i is the row total and C_j is the column total (i=1, 2, ⋯, r; j=1, 2, ⋯, c), N is the grand total ($N = \Sigma R_i = \Sigma C_j = \Sigma\Sigma a_{ij}$), the exclamation symbol (!) represents the *factorial* function (which is the product of all integers from the given number down to 1, e.g., 5!=5×4×3×2×1=120), and a_{ij} is the cell value that corresponds to row i and column j in the matrix.

The *p*-value for Fisher's exact test is the sum of all possible hypergeometric probabilities (conditional on the observed row and column totals) that are less than or equal to the hypergeometric probability of the observed matrix:

[3.4] $$p\text{-value} = \sum(P_h \le P_{ho})$$

where P_{ho} is the hypergeometric probability calculated from [3.3] for the observed matrix.

The concepts and calculations imbedded in [3.3] and [3.4] can be illustrated using the 3×2 matrix in Table 13, taken directly from Table 12 with three species from the ground (G) and inventory (I).

TABLE 13. OBSERVED DATA (LEFT) AND CORRESPONDING VARIABLES (RIGHT) USED TO ILLUSTRATE FISHER'S EXACT TEST.

Species	G	I		Ground (G)	Inventory (I)	Row total
Aw	4	2		a_{11} = 4	a_{12} = 2	$a_{11}+a_{12}$ = 6
Bw	3	6	→	a_{21} = 3	a_{22} = 6	$a_{21}+a_{22}$ = 9
Sb	5	0		a_{31} = 5	a_{32} = 0	$a_{31}+a_{32}$ = 5
Column total	12	8		$a_{11}+a_{21}+a_{31}$ = 12	$a_{12}+a_{22}+a_{32}$ = 8	$N = a_{11}+a_{12}+a_{21}+a_{22}+a_{31}+a_{32}$ = 20

The observed data and the corresponding variables are listed in Table 13 (i=1, 2, 3 and j=1, 2). The observed data have row totals R_1=6, R_2=9 and R_3=5, and column totals C_1=12 and C_2=8. These row and column totals are also known as marginal totals in statistical parlance.

Based on [3.3] and Table 13, the hypergeometric probability of the observed 3×2 matrix is (recognizing 0! = 1):

[3.5] $$P_{ho} = \frac{6!9!5!\times 12!8!}{20!\times(4!2!3!6!5!0!)} = 0.0100024 = 0.0100$$

There are 36 different ways of rearranging the cell frequencies of the observed data in Table 13 into matrices of nonnegative integers while still keeping the marginal totals identical to those from the observed data. These matrices are listed in [3.6]. The matrix that corresponds to the observed cell frequencies is highlighted/shaded in [3.6].

[3.6]
$$\begin{vmatrix}6&0\\1&8\\5&0\end{vmatrix}\ \begin{vmatrix}6&0\\2&7\\4&1\end{vmatrix}\ \begin{vmatrix}6&0\\3&6\\3&2\end{vmatrix}\ \begin{vmatrix}6&0\\4&5\\2&3\end{vmatrix}\ \begin{vmatrix}6&0\\5&4\\1&4\end{vmatrix}\ \begin{vmatrix}6&0\\6&3\\0&5\end{vmatrix}\ \begin{vmatrix}5&1\\2&7\\5&0\end{vmatrix}\ \begin{vmatrix}5&1\\3&6\\4&1\end{vmatrix}\ \begin{vmatrix}5&1\\4&5\\3&2\end{vmatrix}\ \begin{vmatrix}5&1\\5&4\\2&3\end{vmatrix}\ \begin{vmatrix}5&1\\6&3\\1&4\end{vmatrix}\ \begin{vmatrix}5&1\\7&2\\0&5\end{vmatrix}\ \rightarrow$$
$$\begin{vmatrix}4&2\\3&6\\5&0\end{vmatrix}\ \begin{vmatrix}4&2\\4&5\\4&1\end{vmatrix}\ \begin{vmatrix}4&2\\5&4\\3&2\end{vmatrix}\ \begin{vmatrix}4&2\\6&3\\2&3\end{vmatrix}\ \begin{vmatrix}4&2\\7&2\\1&4\end{vmatrix}\ \begin{vmatrix}4&2\\8&1\\0&5\end{vmatrix}\ \begin{vmatrix}3&3\\4&5\\5&0\end{vmatrix}\ \begin{vmatrix}3&3\\5&4\\4&1\end{vmatrix}\ \begin{vmatrix}3&3\\6&3\\3&2\end{vmatrix}\ \begin{vmatrix}3&3\\7&2\\2&3\end{vmatrix}\ \begin{vmatrix}3&3\\8&1\\1&4\end{vmatrix}\ \begin{vmatrix}3&3\\9&0\\0&5\end{vmatrix}\ \rightarrow$$
$$\begin{vmatrix}2&4\\5&4\\5&0\end{vmatrix}\ \begin{vmatrix}2&4\\6&3\\4&1\end{vmatrix}\ \begin{vmatrix}2&4\\7&2\\3&2\end{vmatrix}\ \begin{vmatrix}2&4\\8&1\\2&3\end{vmatrix}\ \begin{vmatrix}2&4\\9&0\\1&4\end{vmatrix}\ \begin{vmatrix}1&5\\6&3\\5&0\end{vmatrix}\ \begin{vmatrix}1&5\\7&2\\4&1\end{vmatrix}\ \begin{vmatrix}1&5\\8&1\\3&2\end{vmatrix}\ \begin{vmatrix}1&5\\9&0\\2&3\end{vmatrix}\ \begin{vmatrix}0&6\\7&2\\5&0\end{vmatrix}\ \begin{vmatrix}0&6\\8&1\\4&1\end{vmatrix}\ \begin{vmatrix}0&6\\9&0\\3&2\end{vmatrix}$$

For each of the 36 matrices in [3.6], the hypergeometric probability is calculated according to [3.3]. For instance, the hypergeometric probabilities of the first (P_{h1}) and second (P_{h2}) matrices in [3.6] are (using four decimal places for final numbers):

[3.7] $$P_{h1} = \frac{6!9!5!\times 12!8!}{20!\times(6!0!1!8!5!0!)} = 0.0000714 = 0.0001 \quad P_{h2} = \frac{6!9!5!\times 12!8!}{20!\times(6!0!2!7!4!1!)} = 0.0014289 = 0.0014$$

Table 14 lists the calculated P_h values for all matrices in [3.6]. The sum of the P_h values for all matrices must equal to one, which is confirmed in the last row of Table 14.

All P_h values less than or equal to the hypergeometric probability of the observed matrix (P_{ho}=0.0100) are listed in the last column in Table 14. Their sum is 0.0643, which is the p-value for Fisher's exact test (see [3.4]). Since this p-value is greater than 0.05, it means that, there is no significant difference between the species proportions (species frequency distributions) from the ground and inventory.

Condition for Using Fisher's Exact Test

Like the chi-square test, Fisher's exact test is used to evaluate the frequency proportions of categorical variables. Unlike the chi-square test, Fisher's exact test does not depend on any large-sample distribution assumptions. It is appropriate even for small sample sizes and for sparse matrices. It can be used irrespective of how small the expected values are.

Fisher's exact test assumes that the row and column totals of an observed matrix are fixed. It uses the hypergeometric distribution to compute the probabilities of all possible matrices of nonnegative integers conditional on the observed row and column totals. Formulating all possible matrices of nonnegative integers conditional on the observed row and column totals of a general r×c matrix can be an extremely exhaustive and arduous process, especially when r and c are three or larger. Even for the simple 3×2 matrix illustrated in Table 13, 35 additional matrices (contingency tables) of nonnegative integers conditional on the observed marginal totals can be formulated. For this reason, Fisher's exact test is typically used only for 2×2 matrices. There appear to be no step-by-step example computations beyond the 2×2 matrices in the limited web search we conducted (this explains why we use a 3×2 matrix to demonstrate the step-by-step computations).

Furthermore, since Fisher's exact test involves the computation of factorials, as in [3.3], and the computed factorials can be astronomically or unimaginably large, the computation burdens associated with Fisher's exact test can be extremely heavy or insurmountable, particularly for medium to large sample sizes of many rows and columns. Even for two innocently looking sample sizes of 20 and 30, the factorials required in computing Fisher's exact test are ginormous (20! = 2,432,902,008,176,640,000 = 2.4329×10^{18} and 30! = 2.6525286×10^{32} - these are quintillions (10^{18}) and decillions (10^{32}) typically appear in astronomy and astrophysics). For this reason, Fisher's exact test may not work for large sample sizes, or it is computationally unmanageable for large sample sizes and large r and c values. The conventional wisdom has been to use Fisher's exact test for small sample sizes only, but there appears to be no consensus in the literature on what constitutes a "small" sample size or a "large" sample size, and where is the quantitative boundary between "small" and "large".

TABLE 14. HYPERGEOMETRIC PROBABILITIES CALCULATED BY [3.3] FOR THE MATRICES LISTED IN [3.6].

Matrix	a_{11}	a_{12}	a_{21}	a_{22}	a_{31}	a_{32}	R_1	R_2	R_3	C_1	C_2	N	P_h	$P_h \le P_{ho}$
1	6	0	1	8	5	0	6	9	5	12	8	20	0.0001	0.0001
2	6	0	2	7	4	1	6	9	5	12	8	20	0.0014	0.0014
3	6	0	3	6	3	2	6	9	5	12	8	20	0.0067	0.0067
4	6	0	4	5	2	3	6	9	5	12	8	20	0.0100	0.0100
5	6	0	5	4	1	4	6	9	5	12	8	20	0.0050	0.0050
6	6	0	6	3	0	5	6	9	5	12	8	20	0.0007	0.0007
7	5	1	2	7	5	0	6	9	5	12	8	20	0.0017	0.0017
8	5	1	3	6	4	1	6	9	5	12	8	20	0.0200	
9	5	1	4	5	3	2	6	9	5	12	8	20	0.0600	
10	5	1	5	4	2	3	6	9	5	12	8	20	0.0600	
11	5	1	6	3	1	4	6	9	5	12	8	20	0.0200	
12	5	1	7	2	0	5	6	9	5	12	8	20	0.0017	0.0017
13	4	2	3	6	5	0	6	9	5	12	8	20	0.0100	P_{ho}=0.0100
14	4	2	4	5	4	1	6	9	5	12	8	20	0.0750	
15	4	2	5	4	3	2	6	9	5	12	8	20	0.1500	
16	4	2	6	3	2	3	6	9	5	12	8	20	0.1000	
17	4	2	7	2	1	4	6	9	5	12	8	20	0.0214	
18	4	2	8	1	0	5	6	9	5	12	8	20	0.0011	0.0011
19	3	3	4	5	5	0	6	9	5	12	8	20	0.0200	
20	3	3	5	4	4	1	6	9	5	12	8	20	0.1000	
21	3	3	6	3	3	2	6	9	5	12	8	20	0.1334	
22	3	3	7	2	2	3	6	9	5	12	8	20	0.0572	
23	3	3	8	1	1	4	6	9	5	12	8	20	0.0071	0.0071
24	3	3	9	0	0	5	6	9	5	12	8	20	0.0002	0.0002
25	2	4	5	4	5	0	6	9	5	12	8	20	0.0150	
26	2	4	6	3	4	1	6	9	5	12	8	20	0.0500	
27	2	4	7	2	3	2	6	9	5	12	8	20	0.0429	
28	2	4	8	1	2	3	6	9	5	12	8	20	0.0107	
29	2	4	9	0	1	4	6	9	5	12	8	20	0.0006	0.0006
30	1	5	6	3	5	0	6	9	5	12	8	20	0.0040	0.0040
31	1	5	7	2	4	1	6	9	5	12	8	20	0.0086	0.0086
32	1	5	8	1	3	2	6	9	5	12	8	20	0.0043	0.0043
33	1	5	9	0	2	3	6	9	5	12	8	20	0.0005	0.0005
34	0	6	7	2	5	0	6	9	5	12	8	20	0.0003	0.0003
35	0	6	8	1	4	1	6	9	5	12	8	20	0.0004	0.0004
36	0	6	9	0	3	2	6	9	5	12	8	20	0.0001	0.0001
Total													**1.0000**	**0.0643**

Note: the observed data (shaded) are listed in Table 13, a_{ij} is the cell value, R_i is the row total and C_j is the column total (i = 1, 2, 3; j = 1, 2), N is the grand total ($N = \Sigma R_i = \Sigma C_j = \Sigma\Sigma a_{ij}$), P_h is the hypergeometric probability defined in [3.3], and P_{ho} is the hypergeometric probability of the observed data. The last column lists the hypergeometric probabilities less than or equal to the hypergeometric probability of the observed data.

Based on numerous tests and evaluations conducted in this study by varying the sample sizes and the row and column numbers of different matrices, it is observed that when the total sample size is greater than about 200 and when $r \times c$ exceeds 16, the computation burdens will most likely become insurmountable in practice. We suggest that Fisher's exact test should only be implemented when more than 20% of the cells in the $r \times c$ matrix have an expected value of less than five. Furthermore, failing to implement Fisher's exact test due to large sample sizes or high dimensionality of a matrix, the chi-square test is a valid approximation and a suitable replacement. The accuracy of the chi-square test increases with the increasing sample sizes. It also increases with the expanding dimensions of the matrix beyond 2×2. An added advantage of the chi-square test is that, computationally, it is much less demanding than Fisher's exact test, regardless of the sample size and the dimensionality of the matrix.

3.4 A note of caution on testing frequency distributions

Since there are some frequent confusions about the chi-square test and Fisher's exact test in the literature, it may be worthwhile to repeat and emphasize here that the chi-square test and Fisher's exact test only evaluate if two species frequency distributions (i.e., species frequency proportions) from two data sets are the same. They do not evaluate if two species frequency counts (i.e., actual numbers) from two data sets are the same.

To elaborate the above statements further, some example data listed in Table 15 are used. The data show the frequency counts in actual numbers and in proportions for the species from the ground and inventory.

TABLE 15. EXAMPLE FREQUENCY (FREQ) COUNTS AND PROPORTIONS FROM THE GROUND AND INVENTORY.

Type		Aw	Bw	Sw	Fb	Lt	Sb	Pb	Pl	Total
Ground	Freq count	1	2	0	1	0	4	3	1	12
	Freq proportion	0.083	0.167	0.000	0.083	0.000	0.333	0.250	0.083	1
Inventory	Freq count	2	1	2	0	1	3	1	0	10
	Freq proportion	0.200	0.100	0.200	0.000	0.100	0.300	0.100	0.000	1

When a vaguely defined term like “species frequency” or “species distribution” is used, it usually conjures up an image of histograms (sometimes also known as frequency bars or frequency charts), similar to those shown in Figure 2 for the data in Table 15.

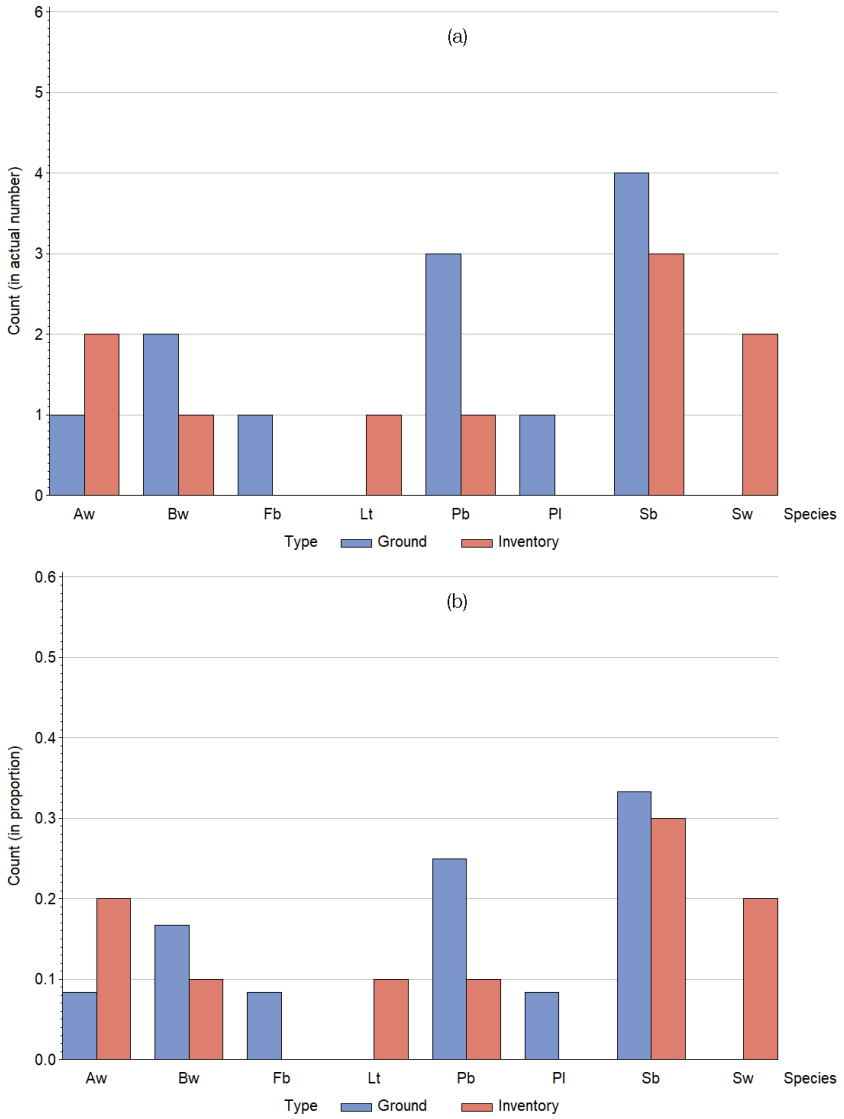

Figure 2. Species frequency counts in actual numbers (top) and in proportions (bottom) from the ground and inventory. Actual data are listed in Table 15.

As mentioned before (in Section 3.1), the word “frequency” could be interpreted to mean frequency counts as in Figure 2(a), or frequency proportions as in Figure 2(b). Many unsuspecting practitioners often thought that the graphs in Figure 2(a) and Figure 2(b) are just two similar ways of displaying the data in Table 15 and either one of them can be used. This is true in the context of showing the data, but not true in the context of the chi-square test and Fisher's exact test, which only apply to Figure 2(b). Both the chi-square test and Fisher's exact test are invariant to the scaling of the data expressed in proportions, but not in actual numbers or counts. Later in Section 5.4, we will provide more details about why this is the case and why the statistical tests only apply to frequency proportions, but not to frequency counts. Practitioners need to remember that the chi-square test and Fisher's exact test for categorical variables (and the Kolmogorov-Smirnov test for continuous variables, to be

discussed later), only enable us to determine if there is a statistically significant difference between two sets of species proportions from two methods, but not two sets of species frequencies in terms of actual counts.

When applying any of the three tests (the chi-square test, Fisher's exact test and the Kolmogorov-Smirnov test), one needs to be careful about the use of the word “distribution”. It is the best to define it clearly and explicitly, as it could mean different things to different people. In statistics and data science, a distribution is typically characterized by three variables: its location (or central tendency) often in terms of the mean, median, or mode; its dispersion (or spread, variability) often in terms of the standard deviation, variance, quartiles, coefficient of variation, range, minimum and maximum; and its shape in terms of the skewness or kurtosis. In day-to-day usages, “distribution” can mean the frequency count or frequency proportion across the size-classes of a continuous variable, or the categories (e.g., different species) of a categorical variable (e.g., species). In this study, frequency distribution is used to explicitly refer to the frequency proportion, not the frequency count, across the size-classes of a continuous variable or the categories of a categorical variable.

Interested readers who want to read more details and see examples about why the chi-square test, Fisher's exact test and the Kolmogorov-Smirnov test only apply to frequency proportions, but not to frequency counts or their statistical distributions are referred to the Additional Notes in Section 5.4.

4 Accuracy and agreement measures for other inventory variables

The methods described so far apply to categorical variables (species and species frequency distribution). The methods described in this Section for the most part apply to continuous variables (focusing on continuous height and density measures in this study). Since remote sensing and photogrammetric studies often involve developing regression models between ground observations and inventory variables (i.e., inventory extracted or derived metrics or measures), the methods for assessing ground = *f* (inventory variables) models are also discussed, together with some cautionary notes on their uses in a regression setting. Readers may benefit from the background and additional technical details described in a preceding study (Huang et al. 2019).

4.1 Goodness-of-fit statistics, agreement measures and plots

Overall Prediction Performance

For continuous variables with continuous numerical values, the following goodness-of-fit statistics are commonly used to measure the overall average performance for all predictions combined. The MAE and RMSE are generally preferred if only two statistics are chosen:

[4.1] $\bar{e} = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i) = \frac{1}{n}\sum e_i$

[4.2] $\text{MAE} = \frac{1}{n}\sum |e_i| = \frac{1}{n}\sum |y_i - x_i|$

[4.3] $\text{RMSE} = \sqrt{\frac{\sum(y_i - x_i)^2}{n}} = \sqrt{s_e^2 + \bar{e}^2}$ (or $\text{MSE} = \frac{\sum(y_i - x_i)^2}{n} = s_e^2 + \bar{e}^2$)

where: $\bar{e}$ is the mean bias (or simply bias); y_i denotes the *i*th observed "true" value or reference value (*i*=1, 2, ⋯, *n*); x_i denotes the *i*th classified, predicted, interpreted, or estimated value (*i*=1, 2, ⋯, *n*); $\sum$ is the summation; e_i is the individual error for the *i*th observation; *n* is the total number of observations; MAE is the mean absolute error; RMSE is the root mean square error and MSE is the mean squared error; and $s_e^2 = \sum(e_i - \bar{e})^2 / n$ is the error variance. Note that the error variance:

$$s_e^2 = \frac{\sum(e_i - \bar{e})^2}{n} = \frac{\sum e_i^2 - \frac{1}{n}[\sum e_i]^2}{n} = \frac{\sum e_i^2 - n\bar{e}^2}{n}.$$

Hence, s_e^2 can also be written as (or rearranged as [4.3]):

$$s_e^2 = \frac{\sum e_i^2}{n} - \bar{e}^2 = \text{RMSE}^2 - \bar{e}^2 = \text{MSE} - \bar{e}^2.$$

The bias $\bar{e}$ describes the deviation of the mean of the predictions (or classifications, interpretations) from the mean of the observed values (considered to be the "truth" and used as the reference). It is caused by the systematic errors in predictions. Since the positive and negative errors in [4.1] can cancel or balance out when summed up (i.e., they may average out to zero or near zero), using the bias alone (and a *t*-test) can sometimes produce "sound good" but misleading results. Therefore, the bias is also expressed in absolute term, as in MAE to alleviate the positive and negative errors' cancelling out problem. The RMSE in [4.3] is the square root of MSE. It measures the total error because it is a combination of both precision (in terms of error variance $\sum(e_i - \bar{e})^2 / n$) and bias (in terms of $\bar{e}^2$). It can be considered an overall accuracy measure.

The statistics expressed in [4.1]-[4.3] can be expressed in percentages in relation to the "truth" or the observed average value ($\bar{y}$):

[4.4] $\bar{e}\% = \frac{100\bar{e}}{\bar{y}}$ $\text{MAE}\% = \frac{100\text{MAE}}{\bar{y}}$ $\text{RMSE}\% = \frac{100\text{RMSE}}{\bar{y}}$

where $\bar{e}$% is the bias percent, MAE% is the mean absolute error percent, RMSE% is percent RMSE or relative RMSE, $\bar{y}$ is the average of the observed reference values ($\bar{y} = \sum y_i / n$) and all other variables are as defined earlier in [4.1]-[4.3].

The $\bar{e}$%, MAE% and RMSE% can be more intuitive in practice than the $\bar{e}$, MAE and RMSE. For instance, sometimes we like to specify that the allowable mean bias ($\bar{e}$) or the mean absolute error (MAE) be within ±10% or ±20% of the observed mean, instead of or in addition to an actual value like ±2 m or ±200 stems.

Individual Prediction Performance

The statistics in [4.1]-[4.4] focus on the overall average performance from all predictions combined, not on individual predictions. Sometimes a more detailed assessment of individual predictions is also needed. It can be very helpful in detecting influential data points and potential outliers, and in knowing how many individual predictions are within certain error limits. For this purpose, the proportions of the observations whose absolute percent errors are smaller than or equal to 10%, 33% (1/3) and 50% (1/2) of the observed values are calculated:

[4.5] $$e_{10} = \frac{\text{Number of } |PE_i| \le 10}{n} \qquad e_{33} = \frac{\text{Number of } |PE_i| \le 33}{n} \qquad e_{50} = \frac{\text{Number of } |PE_i| \le 50}{n}$$

where e_{10}, e_{33} and e_{50} are the proportions of the observations whose absolute percent errors are smaller than or equal to 10%, 33% and 50% of the observed values, respectively, and PE_i is the percent error for the *i*th observation, calculated by:

[4.6] $$PE_i = 100\left(\frac{y_i - x_i}{y_i}\right)$$

where y_i is the *i*th observed value and x_i is its prediction (i=1, 2, ⋯, n). The e_{10}, e_{33} and e_{50} range from 0 (worst) to 1 (best). Larger values of e_{10}, e_{33} and e_{50} indicate better predictions. If necessary, the proportions for other percentages such as e_5, e_{20} and e_{25} (and the maximum and mean of PE_i) can also be calculated following the general expression below, where m can be any meaningful number typically between 1 and 100 and PE_i is defined in [4.6]:

$$e_m = \frac{\text{Number of } |PE_i| \le m}{n} \qquad \text{Mean PE or } \overline{PE} = \frac{1}{n}\sum_{i=1}^{n} PE_i$$

The PE-based statistics can also be used to establish acceptable thresholds in practice. But bear in mind that the allowable error for individual predictions should be much larger (e.g., at least twice or thrice) than that for averages. For instance, we could specify that for a good or an acceptable classification, at least 2/3 of the percent errors should be within 10%, at least half of the percent errors should be within 33%, or at least 95% of the percent errors should be within 50% of the observed values, etc.

Agreement Measure

Among many agreement measures evaluated (Huang et al. 2019), Mielke's measure of agreement (MOA or Mielke's ρ) has several desirable features:

[4.7] $$\text{MOA} = 1 - \frac{\text{MSE}}{\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(y_i - x_j)^2} = 1 - \frac{\text{MSE}}{S_y^2 + S_x^2 + (\bar{y} - \bar{x})^2}$$

where MSE is the mean squared error (RMSE squared), S_y^2 and S_x^2 are variances for y_i and x_i, respectively ($S_y^2 = \sum(y_i - \bar{y})^2 / n$ and $S_x^2 = \sum(x_i - \bar{x})^2 / n$), $\bar{y}$ is the mean of the observed values ($\bar{y} = \sum y_i / n$), $\bar{x}$ is the mean of the predicted values ($\bar{x} = \sum x_i / n$), and all other variables are as defined before.

Calculation of the MOA in [4.7] is fairly straightforward. For example, for three paired *y*-*x* observations, assuming that y = 4, 12, 18 and the corresponding x = 3, 20, 21 (e.g., y = observed on the ground and x = predicted by inventory), the means of the observed and predicted values, the mean squared error and the variances are: $\bar{y} = \frac{\sum y_i}{n}$ = 11.3333, $\bar{x} = \frac{\sum x_i}{n}$ = 14.6667, MSE= $\frac{\sum(y_i - x_i)^2}{n}$ = 24.6667, $S_y^2 = \frac{1}{n}\sum(y_i - \bar{y})^2$ = 32.8889 and $S_x^2 = \frac{1}{n}\sum(x_i - \bar{x})^2$ = 68.2222. Therefore,

$$\text{MOA} = 1 - \frac{MSE}{S_y^2 + S_x^2 + (\bar{y} - \bar{x})^2} = 1 - \frac{24.6667}{32.8889 + 68.2222 + (11.3333 - 14.6667)^2} = 0.7802.$$

The MOA measures the agreement (not correlation!) between y_i and x_i (Mielke 1984, Watterson 1996, Duveiller et al. 2016, Huang et al. 2019). Many other agreement measures (to be discussed later in Section 5.3) are identical, nearly identical or similar to the MOA. The MOA can range from -1 to 1. Larger MOA values indicate better agreement between y_i and x_i. An MOA = 1 implies that all y_i and x_i values, when plotted on the standard *y*-*x* plot, fall on the 45° line that passes through the origin (i.e., a perfect agreement). An MOA = -1 implies that all y_i and x_i values fall on the line that is perpendicular to the 45° line (i.e., a perfect disagreement).

Many additional statistics could also be calculated (e.g., Huang et al. 2013, 2016; Yang and Huang 2014), including standard deviation (SD), coefficient of variation (SD/$\bar{y}$) and relative SD (100SD/$\bar{y}$). But they are often closely related to or do not add much beyond those computed in equations [4.1]-[4.7]. The ubiquitous coefficient of determination (R^2) is more commonly associated with regression analysis and model building, not necessarily model evaluation, accuracy assessment and agreement analysis, although it could be calculated based on the equation defined later in [4.22] for any application data set not used in modeling (yes, it is possible to have an R^2 value < 0 for poor predictions). We will discuss the R^2 in more details

later in Section 4.7. Similarly, some other goodness-of-fit statistics such as Akaike information criterion (AIC), Schwarz's Bayesian information criterion (BIC), Mallows's C_p and predicted residual error sum of squares (PRESS) are typically used only in model building, in the context of variable selection and model selection. They can be applied in area-based approach for variable selection and model selection (i.e., correlation analysis), but they cannot be used to adequately judge the agreement between ground measures and inventory measures (i.e., agreement analysis). Readers are reminded here that correlation analysis and agreement analysis are two very different concepts.

Scatter Plot and Error Plot

In addition to the aforementioned statistics, the agreement between ground measures and inventory predictions should also be assessed through graphical means. Among them, the scatter plot, error plot and Bland-Altman plot are the most informative.

To illustrate, two tree height measurement data sets from a previous study are used (Huang et al. 2019). They are listed in Table 16, where *y* denotes the ground measure (reference) and *x* denotes the corresponding inventory prediction.

TABLE 16. EXAMPLE TREE HEIGHT MEASUREMENT DATA SETS FROM GROUND (*Y*) AND INVENTORY (*X*).

Tree	Data-1					Data-2				
	y	*x*	Diff	Ave	PE	*y*	*x*	Diff	Ave	PE
1	20.2	19.5	0.7	19.85	3.5	18.0	17.1	0.9	17.55	5.0
2	4.0	4.2	-0.2	4.10	-5.0	27.8	26.0	1.8	26.90	6.5
3	14.1	18.5	-4.4	16.30	-31.2	27.4	25.0	2.4	26.20	8.8
4	14.4	19.9	-5.5	17.15	-38.2	29.8	26.4	3.4	28.10	11.4
5	13.6	19.0	-5.4	16.30	-39.7	25.3	24.0	1.3	24.65	5.1
6	27.2	26.5	0.7	26.85	2.6	20.8	20.4	0.4	20.60	1.9
7	7.7	6.6	1.1	7.15	14.3	20.2	18.3	1.9	19.25	9.4
8	25.3	26.9	-1.6	26.10	-6.3	30.0	29.0	1.0	29.50	3.3
9	15.2	15.8	-0.6	15.50	-3.9	35.4	30.9	4.5	33.15	12.7
10	8.1	9.0	-0.9	8.55	-11.1	19.5	19.0	0.5	19.25	2.6
11	25.1	24.9	0.2	25.00	0.8	31.3	27.7	3.6	29.50	11.5
12	19.3	17.0	2.3	18.15	11.9	29.0	22.9	6.1	25.95	21.0
13	19.5	18.7	0.8	19.10	4.1	25.3	24.0	1.3	24.65	5.1
14	23.1	23.4	-0.3	23.25	-1.3	29.3	25.6	3.7	27.45	12.6
15	36.2	30.9	5.3	33.55	14.6	18.1	14.6	3.5	16.35	19.3
16	26.5	28.2	-1.7	27.35	-6.4	26.7	24.7	2.0	25.70	7.5
17	32.4	30.6	1.8	31.50	5.6	27.4	23.0	4.4	25.20	16.1
18	5.4	7.2	-1.8	6.30	-33.3	29.1	23.5	5.6	26.30	19.2
19	16.8	14.8	2.0	15.80	11.9	25.2	22.0	3.2	23.60	12.7
20	33.8	28.7	5.1	31.25	15.1	31.5	24.4	7.1	27.95	22.5
21	23.9	21.3	2.6	22.60	10.9	30.2	24.9	5.3	27.55	17.5
22	21.3	18.7	2.6	20.00	12.2	26.2	20.8	5.4	23.50	20.6
23	16.7	17.0	-0.3	16.85	-1.8	30.2	25.8	4.4	28.00	14.6
24	6.0	7.2	-1.2	6.60	-20.0	23.8	23.1	0.7	23.45	2.9
25	14.3	13.4	0.9	13.85	6.3	20.4	19.7	0.7	20.05	3.4
26	4.6	3.8	0.8	4.20	17.4	27.3	24.0	3.3	25.65	12.1
27	27.9	25.7	2.2	26.80	7.9	20.8	19.8	1.0	20.30	4.8
28	27.5	26.3	1.2	26.90	4.4	32.5	29.0	3.5	30.75	10.8
29	16.6	16.4	0.2	16.50	1.2	25.5	22.8	2.7	24.15	10.6
30	25.8	22.2	3.6	24.00	14.0	21.6	20.1	1.5	20.85	6.9
31	20.8	18.0	2.8	19.40	13.5	21.7	19.5	2.2	20.60	10.1
32	29.1	21.7	7.4	25.40	25.4	22.4	20.2	2.2	21.30	9.8
33	25.1	24.4	0.7	24.75	2.8	19.7	16.3	3.4	18.00	17.3
34	26.3	26.7	-0.4	26.50	-1.5	23.5	22.2	1.3	22.85	5.5
35	29.0	26.6	2.4	27.80	8.3	20.6	17.7	2.9	19.15	14.1
36	8.8	9.1	-0.3	8.95	-3.4	36.0	28.6	7.4	32.30	20.6
37						25.6	22.7	2.9	24.15	11.3
38						30.2	27.6	2.6	28.90	8.6
39						18.7	18.7	0.0	18.70	0.0
40						20.4	19.3	1.1	19.85	5.4
41						23.4	22.2	1.2	22.80	5.1
42						25.6	25.2	0.4	25.40	1.6
Mean	19.77	19.13	0.63		0.14	25.56	22.83	2.73		10.19
SD	8.75	7.66	2.67		15.65	4.72	3.75	1.87		6.08

Note: *y* and *x* represent measures from the ground and inventory, respectively; Diff = *y*–*x*; Ave = (*y*+*x*)/2; PE = 100(*y*-*x*)/*y*; mean refers to the (arithmetic) average; and SD is the standard deviation.

Figure 3 shows five graphs for Data-1 in Table 16 (similar graphs for Data-2 are available but are not shown here). Graph (a) is the standard scatter plot of *y* against *x*. The diagonal line in graph (a) is the line of perfect agreement (i.e., *y*=*x*), also referred to as the line of equality, the line of identity, the unity line, or the 45° line (that passes through the origin). If the *y* and *x* measurements are in good agreement, the data points will be tightly scattered around the 45° line.

The scatter plot is a good starting point, though not sufficient. It only provides a first impression and helps the eye in gauging the degree of agreement between the measurements. However, it usually does not reveal the differences and the trend of the differences well (sometimes it could actually de-sensitize or even "hide" the differences – see error plots next and later). Apparently small differences on the scatter plot may in fact be large. Sometimes the data points may also be clustered near the 45° line and it can be difficult to visually assess the pattern of the differences. In addition, because the eye is better at gauging the departures from a horizontal line than from a slanted line, the error plots (also known as the difference plots or residual plots) are more informative and revealing than the scatter plot.

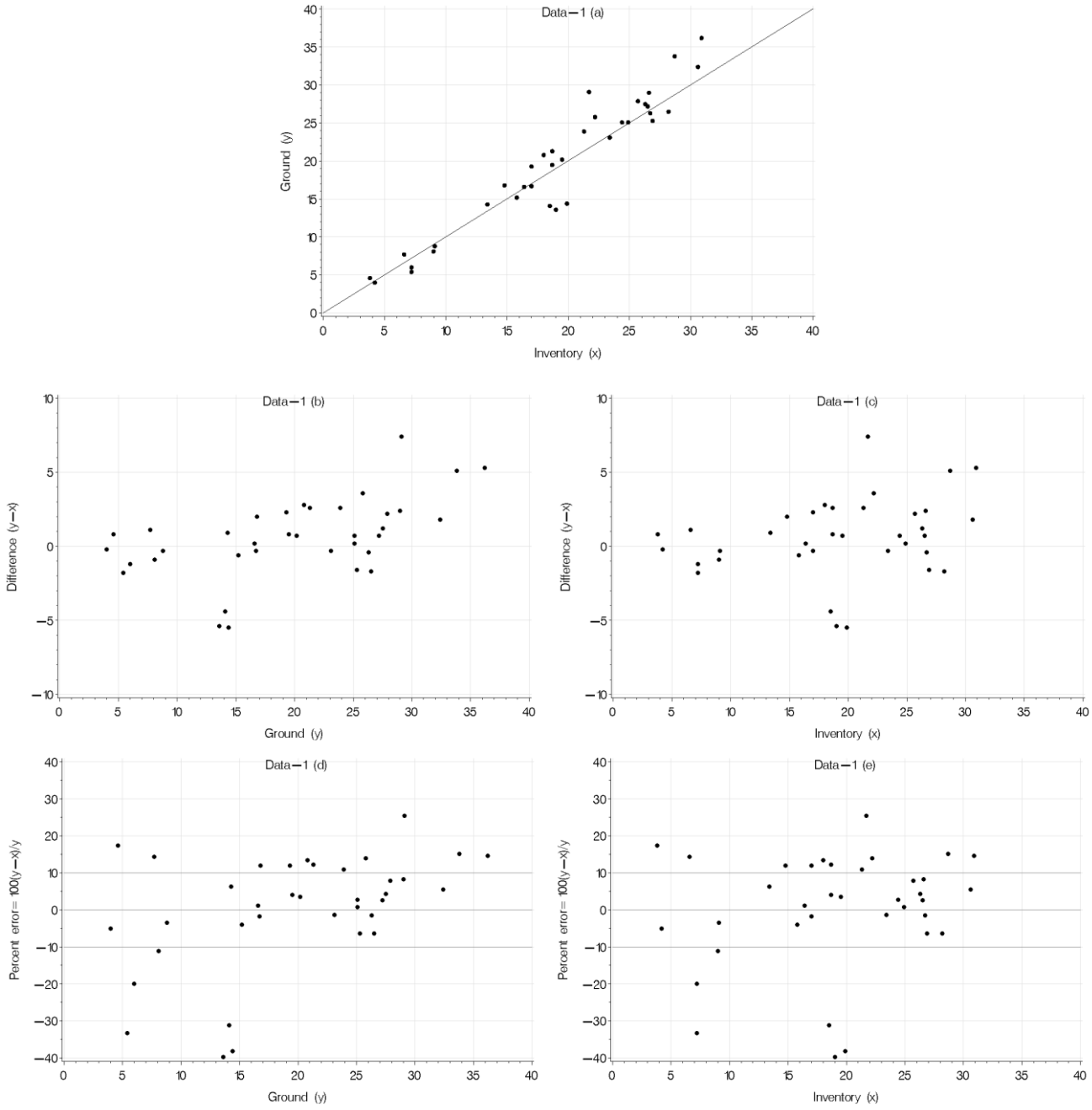

Figure 3. Scatter plot (a) and error plots (b, c, d, e) for Data-1 in Table 16 from the ground (*y*) and inventory (*x*). More detailed descriptions of the plots are provided in the main text.

Graphs (b)-(c) in Figure 3 show the errors in actual values against the ground (*y*) and inventory (*x*) measurements, and graphs (d)-(e) show the errors in percentages (relative to the *y* values). Based on graphs (d)-(e), the observations whose absolute percent errors are greater than 10%, 20%, 30% or any other percentages can be easily seen. The proportions of these observations in relation to the total number of observations can be computed following the expressions given in, e.g., [4.5].

From (b)-(e) it is obvious that some errors are beyond ±5 m and ±30%, which may be considered large. There is also an upward trend seen in graphs (b) and (c), albeit not very strong and obvious. Clearly the error plots (b)-(e) in Figure 3 are more informative than the scatter plot in revealing the differences and the pattern of the differences between *y* and *x*. They align well with the fundamental question of whether the two sets of values from the ground and inventory differ sufficiently small from each other.

In practical accuracy and agreement assessment, the scatter plot of *y* against *x* (graph (a) in Figure 3) needs to be supplemented with at least one of the error plots shown in graphs (b)-(e) in Figure 3, or with the Bland-Altman plot to be discussed next. The scatter plot itself is not sufficient.

The Bland-Altman Plot

Figure 4(a) shows the Bland-Altman plot for Data-1 in Table 16, where the difference (*y*-*x*) is plotted against the average (*y*+*x*)/2. The difference in the Bland-Altman plot can also be expressed in percentage, as in Figure 4(b), where the percent error PE=100(*y*-*x*)/*y* is plotted against the average (*y*+*x*)/2.

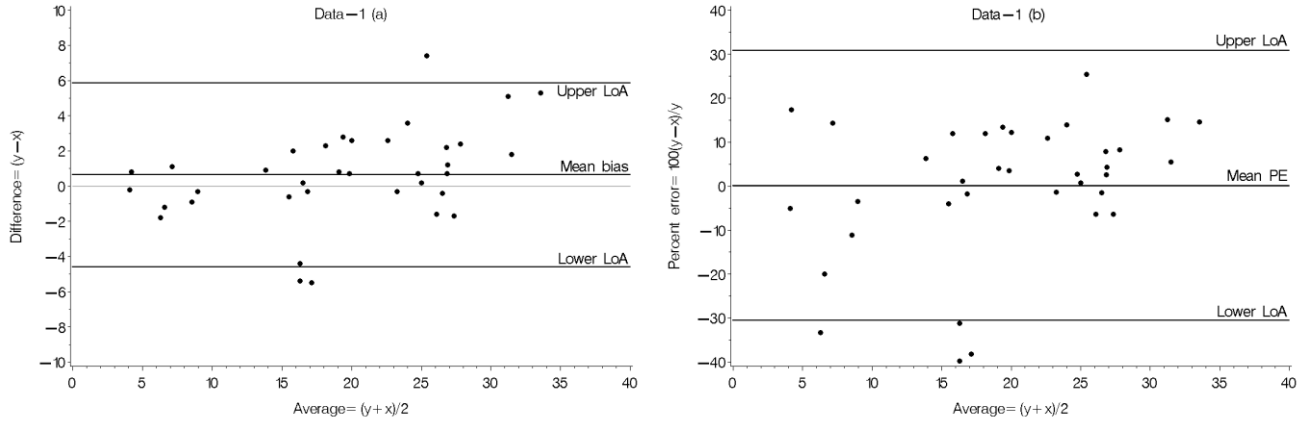

Figure 4. The Bland-Altman plot in actual values (a) and in percentages (b) for Data-1 in Table 16 from the ground (*y*) and inventory (*x*). More detailed descriptions are provided in the main text.

The two solid horizontal lines (labelled “Upper LoA” and “Lower LoA”) in the Bland-Altman plot represent the lower and upper "limits of agreement" (LoAs). In graph (a), the LoAs are defined by ($\bar{e} - 1.96\text{SD}$) and ($\bar{e} + 1.96\text{SD}$), or approximately ($\bar{e} \pm 2\text{SD}$), where $\bar{e}$ is the mean difference (bias) and SD is the standard deviation of the differences. In graph (b), the LoAs are defined by (mean PE – 1.96SD) and (mean PE + 1.96SD), or approximately (mean PE $\pm$ 2SD), where mean PE is the mean percent error and SD is the standard deviation of the percent errors.

Bland and Altman (1986) suggested that one of the considerations for a good agreement between any two sets of measures for a variable is that, 95% or more of the data points in such a plot (originally in actual values only, as in graph (a)) should lie within the lower and upper LoAs:

[4.8] Lower LoA = mean – 1.96SD.

[4.9] Upper LoA = mean + 1.96SD.

where “mean” denotes the mean of the differences and SD is the standard deviation of the differences. For Data-1 in Table 16, since $\bar{e}$=0.633 and SD=2.666, the lower LoA=0.633-(1.96×2.666)=-4.59 and the upper LoA=0.633+(1.96×2.666)=5.86. They define the lower and upper LoA lines in Figure 4(a).

Similarly, for the Bland-Altman plot in percentages, since mean PE=0.14 and SD=15.65, the lower LoA= 0.14-(1.96×15.65)=-30.53 and the upper LoA=0.14+(1.96×15.65)=30.81. They define the lower and upper LoA lines in Figure 4(b).

The Bland-Altman plot (either in actual values or in percentages) can be more intuitive, illuminating and powerful than many calculated statistics and other types of plots (Huang et al. 2019). They can be used to reveal the differences between ground

and inventory measures, detect any abnormalities, identify any systematic trend or bias, and uncover the relationship between the differences and the magnitude of the measurements. They can also be used to pinpoint any possible influential points and outliers, and help establishing reasonable and realistic cut-off thresholds that can be used for accepting or rejecting a new inventory technique.

The Bland-Altman plot alone does not determine the agreement or disagreement between two sets of measurements (e.g., from the ground and inventory, or from inventory techniques 1 and 2). It defines the limits of agreement, but does not say whether these limits are acceptable or not. Acceptable limits are jointly determined by other conditions and individual circumstances, including relevant biological and operational considerations, the inherent variation of the variable in interest, and other subject matter understanding and goals. A more detailed discussion about these conditions and circumstances is provided elsewhere (Huang et al. 2019) and will not be repeated here.

4.2 The Kolmogorov-Smirnov test for continuous variables

For continuous variables like height and density (e.g., stems/ha, actual crown sizes or areas), whether the frequency distribution of the inventory measures is equivalent to that of the ground measures is also a critically important question in determining the validity of an inventory. The Kolmogorov-Smirnov test (KS test for short) is appropriate for answering the question.

The KS test evaluates how well the frequency distributions (i.e., frequency proportions) for two sample data sets, one from the ground and the other from an inventory, conform to each other. The test focuses on the maximum vertical difference (*D*) between the two cumulative distributions (i.e., cumulative proportions) consistent with the two frequency distributions from two samples. Suppose that for variable *y*, the ground samples have a cumulative distribution of $F_G(y)$ and that the inventory samples have a cumulative distribution of $F_I(y)$, the KS test statistic is:

[4.10] $$D = max|F_G(y) - F_I(y)|$$

The null and alternative hypotheses for the KS test are:

H_0: both samples that come from the same population have the same frequency distribution.

H_a: both samples that come from the same population have different frequency distributions.

The critical values for the KS test do not depend on the specific distribution being tested (i.e., the KS test is “distribution-free” in the probability distribution sense). The *p*-value for the KS test is computed by:

[4.11] $$p\text{-value} = 2\sum_{i=1}^{\infty}(-1)^{(i-1)}e^{(-2i^2z^2)} = 2[(-1)^{(1-1)}e^{(-2\times1^2z^2)} + (-1)^{(2-1)}e^{(-2\times2^2z^2)} + (-1)^{(3-1)}e^{(-2\times3^2z^2)} + \cdots + (-1)^{(\infty-1)}e^{(-2\times\infty^2z^2)}]$$

where *e* is the base of natural logarithm or the Euler's number (*e* ≈*2.71828),* and

[4.12] $$z = D\sqrt{\frac{n_1n_2}{n}}$$

where n_1 and n_2 are the sample sizes from two samples and $n=n_1+n_2$.

In theory, calculation of the *p*-value for the KS test involves the summation from *i*=1 to *i*=∞ (infinity). However, after numerous testing on actual and simulated data relevant to our applications, it was observed that any $(-1)^{(i-1)}e^{(-2i^2z^2)}$ value after *i*=3 is 0.00001 or smaller. Hence, for practical purposes, a very accurate approximation for [4.11] can be written as [4.13], which makes the computation much easier in practice (a SAS program that calculates *i*=1 to ∞ is available to interested readers):

[4.13] $$p\text{-value} = 2\sum_{i=1}^{3}(-1)^{(i-1)}e^{(-2i^2z^2)}$$

To illustrate the KS test, some example tree height data from the ground and inventory are used. They are listed in Table 17 and shown in the frequency (left) and cumulative (right) distribution graphs in Figure 5.

TABLE 17. EXAMPLE TREE HEIGHT DATA USED TO ILLUSTRATE THE KOLMOGOROV-SMIRNOV TEST.

Height (HT, m)	12	13	14	15	16	17	18	19	Total	*D*	*p*-value
Ground count	1	2	0	1	0	4	3	1	12	0.2667	0.8327
Inventory count	2	1	2	0	1	3	1	0	10		

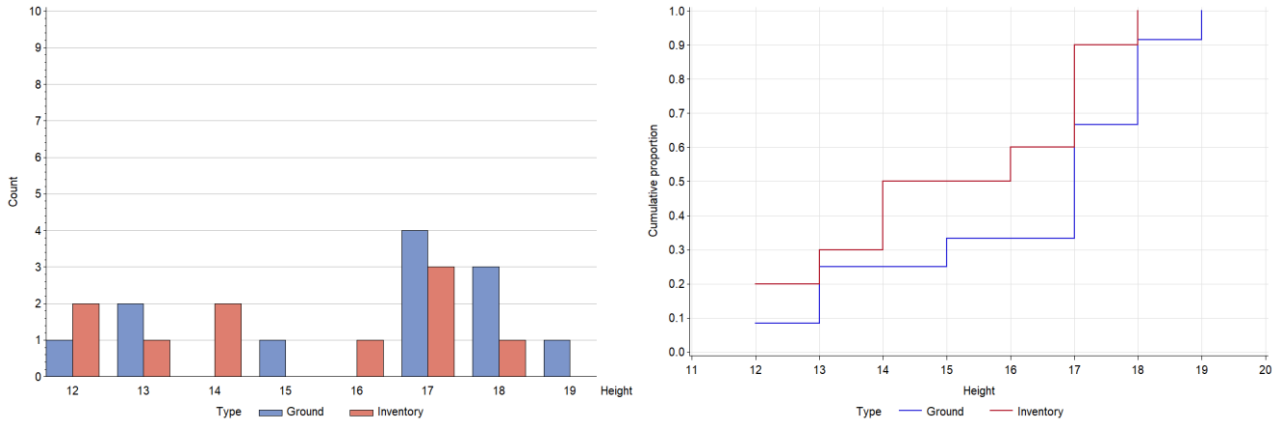

Figure 5. Frequency count (left) and cumulative (right) distribution graphs for the data in Table 17.

Step-by-step computations of the cumulative distributions for the data in Table 17 are shown in Table 18. For instance, for the ground data (n_1=12), each observation (obs) represents 1/12=0.08333 frequency proportion and there are 12 observations, accumulating to a cumulative frequency proportion of 1 (see the first three columns of Table 18). Similar calculation can be done for the inventory data (n_2=10).

TABLE 18. CALCULATING KOLMOGOROV-SMIRNOV TEST STATISTIC FOR THE DATA IN TABLE 17.

Obs	HT	Ground	Obs	HT	Inventory	Obs	HT	Ground	Inventory	*D*
		$F_G(y)$			$F_I(y)$			$F_G(y)$	$F_I(y)$	$\lvert F_G(y) - F_I(y)\rvert$
1	12	0.08333	1	12	0.2	1	12	0.08333	0.2	0.11667
2	13	0.25	2	12	0.2	2	12	0.08333	0.2	0.11667
3	13	0.25	3	13	0.3	3	13	0.25	0.3	0.05
4	15	0.33333	4	14	0.5	4	13	0.25	0.3	0.05
5	17	0.66667	5	14	0.5	5	14	0.25	0.5	0.25
6	17	0.66667	6	16	0.6	6	14	0.25	0.5	0.25
7	17	0.66667	7	17	0.9	7	15	0.33333	0.5	0.16667
8	17	0.66667	8	17	0.9	8	**16**	**0.33333**	**0.6**	**0.26667**
9	18	0.91667	9	17	0.9	9	17	0.66667	0.9	0.23333
10	18	0.91667	10	18	1	10	17	0.66667	0.9	0.23333
11	18	0.91667				11	17	0.66667	0.9	0.23333
12	19	1				12	17	0.66667	0.9	0.23333
n_1 = 12			n_2 = 10			13	18	0.91667	1	0.08333
						14	18	0.91667	1	0.08333
						15	18	0.91667	1	0.08333
						16	19	1	1	0

Note: $F_G(y)$ and $F_I(y)$ are cumulative distributions for the ground (G) and inventory (I) data, respectively. Actual data are listed in Table 17.
In order to compute the vertical difference between the two cumulative distributions from the ground and inventory, the cumulative distributions from the ground and inventory are arranged into a format shown in the right-half of Table 18 (through simple ranking). The distances between $F_G(y)$ and $F_I(y)$ are listed in the last column of Table 18. The KS test statistic (D), which is the maximum (absolute) distance between $F_G(y)$ and $F_I(y)$, is 0.26667, which occurs at a height of 16 (m), highlighted in Table 18.

With the known D=0.26667, the z is calculated according to [4.12]:

$$z = D\sqrt{\frac{n_1 n_2}{n}} = 0.26667\sqrt{\frac{12\times10}{22}} = 0.62280.$$

Therefore, the p-value for the KS test can be computed by [4.11] or [4.13]:

$$p\text{-value} = 2[(-1)^{(1-1)}e^{(-2z^2)} + (-1)^{(2-1)}e^{(-2\times2^2z^2)} + (-1)^{(3-1)}e^{(-2\times3^2z^2)}] + \cdots$$

$$= 2[0.46035 + (-0.04491) + 0.00093 + (-0.00000) + 0.00000 + \cdots] = 0.8327.$$

This p-value is greater than α=0.05, which means that the null hypothesis cannot be rejected. It suggests that the samples from the ground and inventory follow the same frequency distribution.

To demonstrate the KS test further using actual values (instead of the counts by height classes as in Table 17), the tree height data listed in Table 16 (Data-1 and Data-2) are used. To limit the size of the upcoming table, only the first 18 observations

from Data-1 are used in the step-by-step computations demonstrated below. Figure 6 shows the cumulative distributions for the first 18 trees of Data-1 (left), and for all trees of Data-2 (right). Since every one of the 18 tree heights from Data-1 is unique, the frequency distribution for Data-1 is uniform (i.e., the frequency count that corresponds to each tree height is 1).

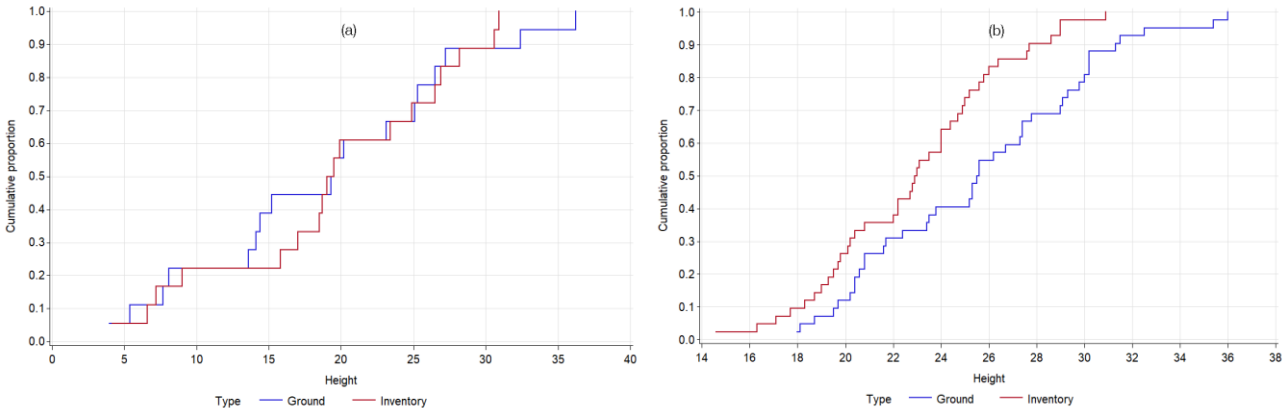

Figure 6. Cumulative distributions for the first 18 tree heights of Data-1 (left), and for all tree heights of Data-2 (right). Data-1 and Data-2 are listed in Table 16.

Table 19 shows the cumulative frequency calculations and the data arrangement required for computing the maximum vertical difference (the KS test statistic *D*) between the two cumulative distributions from the ground and inventory. The logic is identical to that illustrated in Table 18.

TABLE 19. CALCULATING KOLMOGOROV-SMIRNOV TEST STATISTIC FOR DATA-1 (FIRST 18 OBSERVATIONS) IN TABLE 16.

Obs	HT	Ground $F_G(y)$	HT	Inventory $F_I(y)$	Obs	HT	Ground $F_G(y)$	Inventory $F_I(y)$	D $\lvert F_G(y) - F_I(y)\rvert$
1	4.0	0.05556	4.2	0.05556	1	4.0	0.05556	0.00000	0.05556
2	5.4	0.11111	6.6	0.11111	2	4.2	0.05556	0.05556	0.00000
3	7.7	0.16667	7.2	0.16667	3	5.4	0.11111	0.05556	0.05556
4	8.1	0.22222	9.0	0.22222	4	6.6	0.11111	0.11111	0.00000
5	13.6	0.27778	15.8	0.27778	5	7.2	0.11111	0.16667	0.05556
6	14.1	0.33333	17.0	0.33333	6	7.7	0.16667	0.16667	0.00000
7	14.4	0.38889	18.5	0.38889	7	8.1	0.22222	0.16667	0.05556
8	15.2	0.44444	18.7	0.44444	8	9.0	0.22222	0.22222	0.00000
9	19.3	0.50000	19.0	0.50000	9	13.6	0.27778	0.22222	0.05556
10	19.5	0.55556	19.5	0.55556	10	14.1	0.33333	0.22222	0.11111
11	20.2	0.61111	19.9	0.61111	11	14.4	0.38889	0.22222	0.16667
12	23.1	0.66667	23.4	0.66667	12	**15.2**	**0.44444**	**0.22222**	**0.22222**
13	25.1	0.72222	24.9	0.72222	13	15.8	0.44444	0.27778	0.16667
14	25.3	0.77778	26.5	0.77778	14	17.0	0.44444	0.33333	0.11111
15	26.5	0.83333	26.9	0.83333	15	18.5	0.44444	0.38889	0.05556
16	27.2	0.88889	28.2	0.88889	16	18.7	0.44444	0.44444	0.00000
17	32.4	0.94444	30.6	0.94444	17	19.0	0.44444	0.50000	0.05556
18	36.2	1.00000	30.9	1.00000	18	19.3	0.50000	0.50000	0.00000
n_1 = 18			n_2 = 18		19	19.5	0.55556	0.55556	0.00000
					20	19.9	0.55556	0.61111	0.05556
					21	20.2	0.61111	0.61111	0.00000
					22	23.1	0.66667	0.61111	0.05556
					23	23.4	0.66667	0.66667	0.00000
					24	24.9	0.66667	0.72222	0.05556
					25	25.1	0.72222	0.72222	0.00000
					26	25.3	0.77778	0.72222	0.05556
					27	26.5	0.83333	0.77778	0.05556
					28	26.9	0.83333	0.83333	0.00000
					29	27.2	0.88889	0.83333	0.05556
					30	28.2	0.88889	0.88889	0.00000
					31	30.6	0.88889	0.94444	0.05556
					32	30.9	0.88889	1.00000	0.11111
					33	32.4	0.94444	1.00000	0.05556
					34	36.2	1.00000	1.00000	0.00000

Note: $F_G(y)$ and $F_I(y)$ are cumulative distributions for the ground (G) and inventory (I) data, respectively.

Based on the $F_G(y)$ and $F_I(y)$ values in Table 19, the maximum (absolute) distance between $F_G(y)$ and $F_I(y)$ is calculated to be D=0.22222, which occurs at a height of 15.2 m. With the known D, the z is calculated according to [4.12]:

$$z = D\sqrt{\frac{n_1 n_2}{n}} = 0.22222\sqrt{\frac{18\times 18}{36}} = 0.66667.$$

Having the z value, the p-value can be computed by [4.11] or [4.13]:

$$p\text{-value} = 2[(-1)^{(1-1)}e^{(-2z^2)} + (-1)^{(2-1)}e^{(-2\times 2^2 z^2)} + (-1)^{(3-1)}e^{(-2\times 3^2 z^2)}] + \cdots$$

$$= 2[0.41111 + (-0.02857) + 0.00034 + (-0.00000) + 0.00000 + \cdots] = 0.7658.$$

This p-value is greater than α=0.05, suggesting that the first 18 tree heights of Data-1 from the ground and inventory follow the same frequency distribution.

The KS test can be used to assess whether the frequency distributions from any two sets of samples are the same or different. Following the same steps demonstrated above, the KS test statistic and the p-value for Data-2 in Table 16 are computed (a step-by-step SAS program is available to interested readers, and a generalized program that contains only four keywords is also available to practitioners who may not want to know the step-by-step details). Results are listed here for interested readers who may wish to check: D=0.3333 (which occurs at HT=25.0 and HT=25.2) and p-value=0.0188. This p-value is smaller than 0.05, suggesting that the tree heights from the ground and inventory for Data-2 follow different frequency distributions. Indeed, the difference in the frequency distributions for Data-2 can be seen clearly from the cumulative distribution graph shown in Figure 6(b).

4.3 Tree height

Felled Tree Height vs Lidar (Inventory) Height

To further demonstrate the application of the methods described in Sections 4.1 and 4.2, a total of 108 trees of various species and sizes were located using a global navigation satellite system (GNSS) within the Forest Management Agreement (FMA) area of Canadian Forest Products Ltd. (Grande Prairie). Their heights were measured/extracted from the high density (>16 pts/m^2) airborne lidar. These trees were cut down and their heights were also measured on the ground using a measuring tape. Table 20 lists the tree heights measured on the ground (HT_{felled}) and from the lidar (HT_{lidar}).

TABLE 20 (PART 1 OF 2). TREE HEIGHTS (M) MEASURED ON THE GROUND (HT_{FELLED}) AND FROM LIDAR (HT_{LIDAR}).

Tree	Sp	HT_{felled}	HT_{lidar}	Tree	Sp	HT_{felled}	HT_{lidar}	Tree	Sp	HT_{felled}	HT_{lidar}
1	Sw	25.60	25.75	37	Pl	23.54	23.07	73	Sw	20.77	20.30
2	Aw	23.95	23.97	38	Dp	20.06	21.71	74	Sw	6.23	5.97
3	Pb	18.01	20.97	39	Pl	13.62	12.97	75	Sw	15.60	18.40
4	Aw	18.80	22.07	40	Pl	21.43	21.40	76	Sw	16.17	18.07
5	Aw	19.25	19.54	41	Pl	20.22	20.14	77	Sw	12.50	15.97
6	Sw	14.08	12.14	42	Pl	21.91	21.63	78	Sw	11.37	9.07
7	Dp	19.45	20.78	43	Pl	21.12	20.76	79	Aw	18.97	19.56
8	Sw	18.96	17.91	44	Pl	21.78	21.57	80	Aw	9.30	14.96
9	Sw	19.86	19.80	45	Sw	10.18	10.23	81	Aw	18.90	18.96
10	Aw	19.65	19.35	46	Sw	18.97	17.98	82	Aw	19.17	18.18
11	Sw	11.87	11.77	47	Sw	20.70	19.87	83	Aw	20.02	18.41
12	Aw	23.42	23.56	48	Sw	13.45	16.84	84	Aw	10.00	12.68
13	Aw	22.18	23.30	49	Sw	17.95	20.71	85	Aw	15.76	16.21
14	Pl	16.66	18.88	50	Sw	16.12	15.07	86	Aw	17.64	18.85
15	Pl	19.38	19.03	51	Sw	15.20	14.76	87	Aw	18.91	19.10
16	Sw	8.66	8.48	52	Sw	16.09	15.60	88	Aw	19.71	20.63
17	Aw	20.60	18.23	53	Sw	21.11	21.28	89	Aw	17.49	18.44
18	Sw	12.38	13.39	54	Sw	7.42	6.76	90	Aw	10.91	10.88
19	Aw	17.14	16.39	55	Aw	21.35	21.13	91	Aw	16.72	18.81
20	Aw	21.71	20.37	56	Aw	20.06	20.37	92	Aw	22.89	20.91
21	Sw	16.22	19.50	57	Aw	20.00	19.36	93	Aw	12.57	12.58
22	Aw	24.23	26.18	58	Sw	16.38	15.93	94	Aw	15.38	17.61
23	Aw	25.15	25.06	59	Sw	14.97	15.25	95	Aw	11.04	10.63
24	Aw	25.04	23.85	60	Sw	17.19	16.78	96	Aw	19.95	19.50
25	Aw	26.32	25.18	61	Sw	20.40	20.19	97	Aw	18.35	18.29

Note: tree species (sp) are defined in Table 1. HT_{felled} and HT_{lidar} denote felled tree height (m) and lidar height (m), respectively.

TABLE 20 (PART 2 OF 2). TREE HEIGHTS (M) MEASURED ON THE GROUND (HT_{FELLED}) AND FROM LIDAR (HT_{LIDAR}).

Tree	Sp	HT_{felled}	HT_{lidar}	Tree	Sp	HT_{felled}	HT_{lidar}	Tree	Sp	HT_{felled}	HT_{lidar}
26	Sw	14.82	13.87	62	Sw	9.34	9.28	98	Aw	18.62	20.62
27	Sw	15.60	16.26	63	Sw	24.18	23.36	99	Aw	10.18	9.77
28	Aw	22.54	22.34	64	Sw	19.46	19.09	100	Sw	6.26	5.12
29	Sw	12.87	13.94	65	Sw	11.42	13.50	101	Sw	10.92	12.99
30	Aw	21.36	22.10	66	Sw	14.99	17.91	102	Sw	10.66	12.91
31	Sw	10.68	11.37	67	Aw	23.60	23.51	103	Sw	9.56	8.18
32	Sw	11.22	10.46	68	Sw	16.30	14.98	104	Sw	6.77	6.17
33	Sw	22.91	22.12	69	Sw	17.37	19.13	105	Sw	6.35	8.26
34	Bw	17.50	21.63	70	Sw	23.33	23.65	106	Aw	6.61	8.60
35	Sw	11.68	10.88	71	Sw	24.08	22.66	107	Aw	8.16	6.90
36	Sw	24.62	24.25	72	Sw	14.31	14.07	108	Aw	8.01	12.84

Note: tree species (sp) are defined in Table 1. HT_{felled} and HT_{lidar} denote felled tree height (m) and lidar height (m), respectively.

Figure 7 shows the scatter plot of felled heights against lidar heights, along with two error plots where the percent error in (c) is calculated as $100(HT_{felled} - HT_{lidar})/HT_{felled}$. It can be seen from the scatter plot that the data points are (more or less) scattered fairly tightly around the 45° line – an indication of good agreement between HT_{felled} and HT_{lidar}. The error plots in (b) and (c) show that there are several data points whose errors are around -5 m or more than -30%. They also show that the errors appear to be unequally varied across the range of lidar heights (this is not immediately clear from the scatter plot). For some analysis this could invoke some specific methods capable of handling the unequal variation.

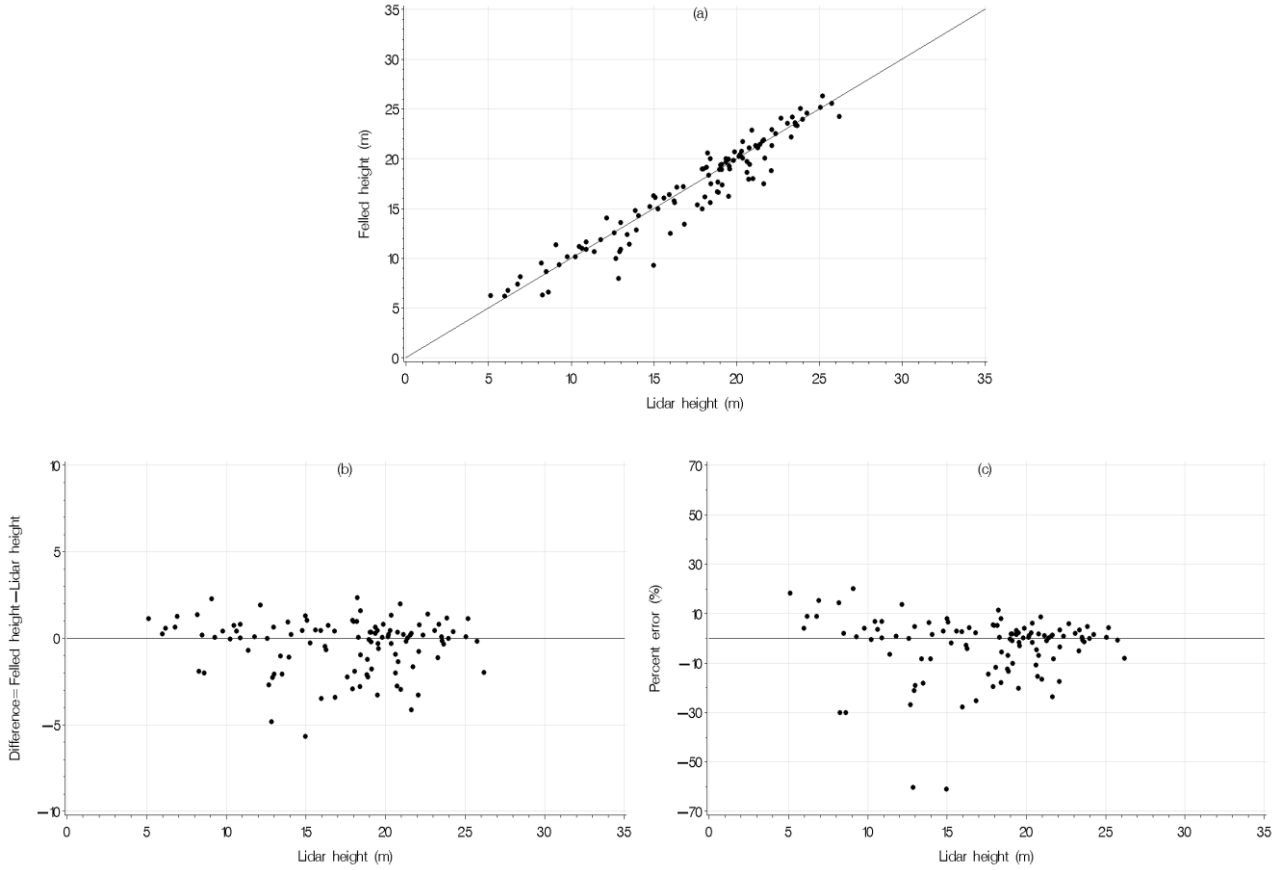

Figure 7. Scatter plot (a) and error plots (b, c) for felled tree heights (HT_{felled}) and lidar heights (HT_{lidar}). Actual data are listed in Table 20. The percent error in (c) is calculated as $100(HT_{felled} - HT_{lidar})/HT_{felled}$.

Table 21 lists the goodness-of-fit statistics and Mielke's measure of agreement. The statistics and MOA are also calculated by the broad cover types (coniferous and deciduous) defined in Table 1. It can be seen from Table 21 that the lidar measurements for coniferous are more accurate than those for deciduous. This makes biological sense because the deciduous species usually have more large branches and multiple leaders (and irregular crown shapes), which can make stem segmentation and measurement from lidar point clouds more difficult. Overall, the errors are small (e.g., $\bar{e}$%=-2.2%) and the

agreement between felled and lidar heights is quite strong (MOA=0.952). More than 2/3 of the percent errors are within 10% (e_{10}=0.731) and more than 95% of the percent errors are within 33% (e_{33}=0.981).

TABLE 21. GOODNESS-OF-FIT STATISTICS AND AGREEMENT MEASURE BETWEEN FELLED AND LIDAR HEIGHTS.

Type	*n*	$\bar{e}$	MAE	RMSE	$\bar{e}$%	MAE%	RMSE%	e_{10}	e_{33}	e_{50}	MOA
All	108	-0.371	1.128	1.577	-2.2%	6.7%	9.3%	0.731	0.981	0.981	0.952
Coniferous	63	-0.236	1.040	1.390	-1.5%	6.5%	8.7%	0.730	1	1	0.963
Deciduous	45	-0.561	1.251	1.807	-3.1%	6.9%	10.0%	0.733	0.956	0.956	0.928

Note: *n* denotes the sample size, $\bar{e}$, MAE, RMSE, $\bar{e}$%, MAE%, RMSE%, e_{10}, e_{33}, e_{50} and MOA are defined in equations [4.1]-[4.5] and [4.7].

To assess the agreement further between felled and lidar heights, the Bland-Altman plots in actual values and in percentages are shown in Figure 8.

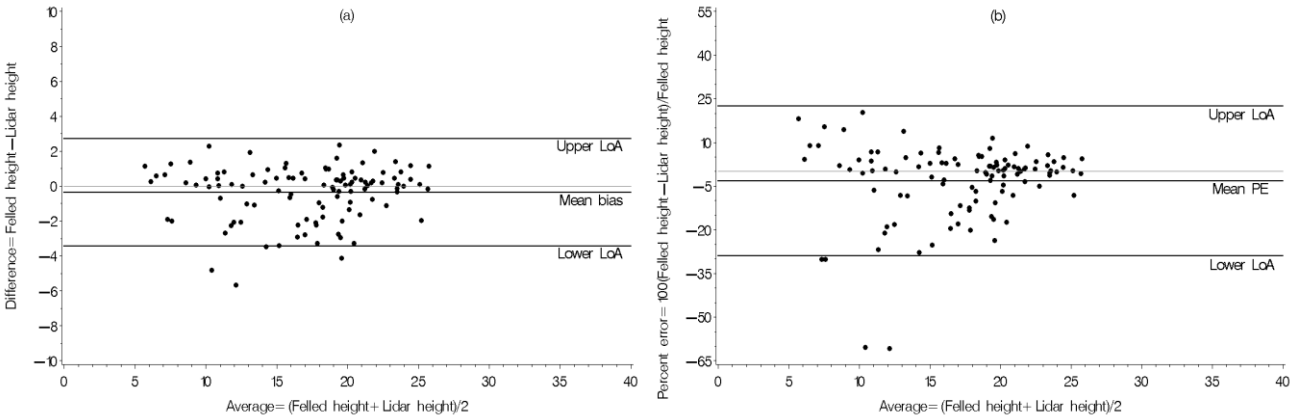

Figure 8. The Bland-Altman plots in actual values (a) and in percentages (b). More detailed descriptions of the plots are provided in the main text.

In Figure 8(a), the differences (HT_{felled} - HT_{lidar}) are plotted against their averages (HT_{felled} + HT_{lidar})/2. The mean bias is $\bar{e}$=-0.371 and the standard deviation of the differences is SD=1.540. Therefore, the Lower LoA=-0.371-2×1.540=-3.451 and the Upper LoA=-0.371+2×1.540=2.709.

In Figure 8(b), the percent errors (100(HT_{felled} - HT_{lidar})/HT_{felled}) are plotted against the averages (HT_{felled} + HT_{lidar})/2. The mean percent error is $\overline{\text{PE}}$%=-3.159, and the standard deviation of the percent errors is SD=12.865. Therefore, the Lower LoA=-3.159-2×12.865=-28.889 and the Upper LoA=-3.159+2×12.865= 22.571. The percent errors are generally larger for shorter trees, reflecting the fact that stem segmentation and measurement from lidar point clouds can be more difficult and varied for shorted trees.

In both graphs in Figure 8, only four data points are outside the LoA lines, implying that (108-4)/108= 96% of the data points lie within the lower and upper LoAs. This, together with Figure 7 and Table 21, is an indication of a good agreement between felled and lidar heights. If necessary, the data points with large errors or abnormalities can be pinpointed easily based on the error plots or Bland-Altman plots.

The Kolmogorov-Smirnov Test between Felled and Lidar Heights

Following the KS test described earlier (Section 4.2), the two cumulative distribution functions from felled heights and lidar heights are shown in Figure 9.

The maximum (absolute) distance between the two cumulative distributions from felled and lidar heights is calculated to be *D*=0.092593, which occurs at the heights of 17.50 and 17.64 (m). With the known *D*, the *z* is calculated according to [4.12]: $z = 0.092593\sqrt{108 \times 108/216}$ = 0.68041. Hence, the *p*-value can be computed by [4.11]: *p*-value = 2[0.39616+ (-0.02463) + 0.00024 + (-0.00000) + 0.00000 + ⋯] = 0.7435. This *p*-value is greater than α=0.05, suggesting that felled heights and lidar heights follow the same frequency distribution.

Based on all of the above assessments (from Figures 7-8, Table 21 and the KS test), it can be inferred that for this data set, the agreement between the tree heights measured on the ground and from lidar point clouds is reasonably good. Lidar heights are representative of the ground heights and can be used to substitute the ground heights in general, although further improvement to lidar heights may still be possible through additional adjustment or calibration (see Section 5.6).

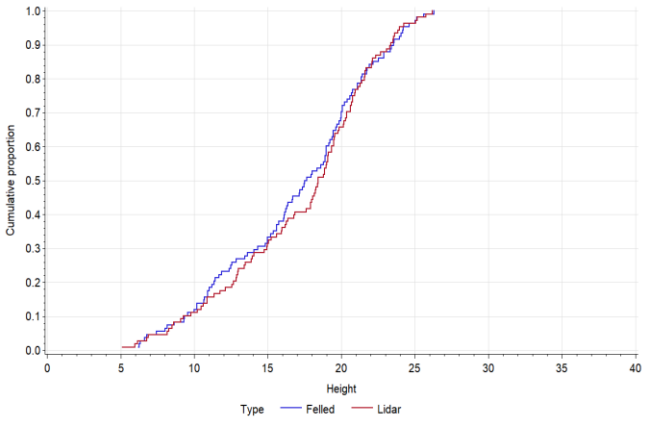

Figure 9. Cumulative distributions from felled and lidar heights. Actual data are listed in Table 20.

4.4 Stand height

For stand-level height measurements obtained on the ground and from any inventory technique, similar comparison can be done following the same procedures demonstrated above for individual tree height measurements. The only caveat about stand height comparison is that, "stand height" must be clearly defined and consistently used. Otherwise, we could be comparing apples and lemons.

Historically, the concept of stand height is not precise, nor consistent. Different stand heights have been used in different studies. This is illustrated in Table 22, where the data from 12 felled trees (assumed to be from a 100 m^2 plot) are listed and used to derive different stand heights. These 12 felled trees are the first 12 trees listed in Table 20, each with two additional variables (DBH and crown position) measured.

TABLE 22. TREE DATA FROM A 100 M^2 PLOT USED TO CALCULATE STAND HEIGHTS.

Tree	Sp	DBH (cm)	Crown position	Height (m)	Freq	Percent	Cum freq	Cum percent	Top 30%	Top 20%	Top 10%	Top 5%
					Ranked by Height							
11	Sw	14.9	U	11.87	1	8.33	1	8.33	.	.	.	.
6	Sw	15.0	U	14.08	1	8.33	2	16.67	.	.	.	.
3	Pb	17.3	U	18.01	1	8.33	3	25.00	.	.	.	.
4	Aw	17.7	O	18.80	1	8.33	4	33.33	.	.	.	.
8	Sw	19.6	O	18.96	1	8.33	5	41.67	.	.	.	.
5	Aw	23.4	O	19.25	1	8.33	6	50.00	.	.	.	.
7	Dp	22.8	O	19.45	1	8.33	7	58.33	.	.	.	.
10	Aw	19.8	O	19.65	1	8.33	8	66.67	.	.	.	.
9	Sw	20.9	O	19.86	1	8.33	9	75.00	19.86	.	.	.
12	Aw	25.0	O	23.42	1	8.33	10	83.33	23.42	23.42	.	.
2	Aw	37.7	O	23.95	1	8.33	11	91.67	23.95	23.95	23.95	.
1	Sw	35.0	O	25.60	1	8.33	12	100.00	25.60	25.60	25.60	25.60
					Ranked by DBH							
11	Sw	14.9	U	11.87	1	8.33	1	8.33	.	.	.	.
6	Sw	15.0	U	14.08	1	8.33	2	16.67	.	.	.	.
3	Pb	17.3	U	18.01	1	8.33	3	25.00	.	.	.	.
4	Aw	17.7	O	18.80	1	8.33	4	33.33	.	.	.	.
8	Sw	19.6	O	18.96	1	8.33	5	41.67	.	.	.	.
10	Aw	19.8	O	19.65	1	8.33	6	50.00	.	.	.	.
9	Sw	20.9	O	19.86	1	8.33	7	58.33	.	.	.	.
7	Dp	22.8	O	19.45	1	8.33	8	66.67	.	.	.	.
5	Aw	23.4	O	19.25	1	8.33	9	75.00	19.25	.	.	.
12	Aw	25.0	O	23.42	1	8.33	10	83.33	23.42	23.42	.	.
1	Sw	35.0	O	25.60	1	8.33	11	91.67	25.60	25.60	25.60	.
2	Aw	37.7	O	23.95	1	8.33	12	100.00	23.95	23.95	23.95	23.95

Note: species (sp) are defined in Table 1, crown position refers to the understory (U) or overstory (O) assigned in the field by the surveyor, freq denotes frequency count, cum freq and cum percent are cumulative frequency count and cumulative percent, respectively, top 30%, top 20%, top 10% and top 5% refer to the tallest (ranked by height) or largest (ranked by DBH) 30%, 20%, 10% and 5% of the trees in the plot.

Before demonstrating the calculation of different stand heights, understanding the concept of a ceiling function is helpful, which is used in selecting a set of trees from a tree list for stand height calculation.

When calculating the number of trees that fall within the top percentages of the trees from a tree list, we usually round up, by taking the "smallest integer that is greater than or equal to the real number" (i.e., we usually take the "ceiling function" of the real number). For instance, when selecting 5, 10, 20 and 30% of the trees from a total of 12 trees listed in Table 22:

5% of the trees = 12×5% = 0.6 = 1 tree.

10% of the trees = 12×10% = 1.2 = 2 trees.

20% of the trees = 12×20% = 2.4 = 3 trees.

30% of the trees = 12×30% = 3.6 = 4 trees.

Where 0.6, 1.2, 2.4 and 3.6 are the real numbers and 1, 2, 3 and 4 are the smallest integers that are greater than or equal to the real numbers, respectively (the standard rounding of 0.6, 1.2, 2.4 and 3.6 would be 1, 1, 2 and 4, respectively). Using a ceiling function (or a "floor function", i.e., rounding down) or the standard rounding can influence the result considerably when the sample size is not big.

The following stand heights are defined and calculated based on the data in Table 22. Here the calculations are done for all species combined. When necessary and relevant (e.g., for mixed-species stands and/or for multi-cohort stand structures), they can also be done by species, species groups, layers or cohorts, or any other user-defined strata.

1. **Average height** (H_{ave}) – generally refers to the (arithmetic) average height of all trees in a stand. For the data in Table 22, H_{ave}=19.41 (m).

   Often, since it is unlikely that "all trees" in a stand will be measured for heights, average height in practice usually means the average height of a portion of trees in a stand. For instance, the average height of all trees taller than "1.3, 2, 5 or 10 m in height", or the average height of all trees larger than "5.0 or 9.0 cm in DBH". In a more generic sense, average height can mean the average height of any user-defined set of trees or threshold in a stand.

2. **Dominant and co-dominant height** (H_{dom}) – refers to the average height of the dominant and co-dominant trees in a stand. Sometimes, dominant and co-dominant height is also referred to as "overstory height". Since the definition of what is "dominant" or "co-dominant" can be imprecise and ambiguous in operation (e.g., calling a tree "dominant", "co-dominant" or "intermediate" can sometimes be quite arbitrary depending on the stand type, stand structure and the specific location a surveyor is standing in the stand), and the number of "dominants" and "co-dominants" required per unit area to derive H_{dom} can be arbitrary as well (e.g., sometimes we require one dominant and two co-dominants, or some other combinations of dominants and co-dominants), the use of this stand height requires caution (as it often lacks consistency). For the data in Table 22 assigned with a crown position of "O", H_{dom}=20.99 (m).

3. **Top height, tree height-ranked** (H_{top}) – refers to the average height of the 100 tallest trees per hectare. For the data in Table 22, H_{top}=25.60 (m).

4. **Top percent height, tree height-ranked** (H%) – refers to the average height of the tallest 5%, 10%, 20% or 30% of the trees per unit area. Other percentages (e.g., 25%, 33%, 50%, 66%) may also be used (especially when studying stand structures). For the data in Table 22:

   H5 (tallest 5%) = 25.60 (m).
   H10 (tallest 10%) = (23.95 +25.60)/2 = 24.78 (m).
   H20 (tallest 20%) = (23.42+23.95 +25.60)/3 = 24.32 (m).
   H30 (tallest 30%) = (19.86+23.42+23.95 +25.60)/4 = 23.21 (m).

   The top percent height can be considered complementary to the percentile height. In statistics, a percentile is often defined as a value below which a given percentage of all values in its frequency distribution falls. Hence, the average height of the top 5%, 10%, 20% or 30% trees corresponds to the average height of all trees at and above the 95th, 90th, 80th or 70th percentile, respectively.

   Among the potential top percent heights, if only one single top percent height is needed, the average height of the tallest 20% or 25% of the trees per unit area is generally preferred. The tallest 20% of the trees correspond well in most cases with the traditional stand height (i.e., the average height of the dominant and co-dominant trees) commonly used in Alberta, and it is quantified thus more precise, consistent and repeatable than the vaguely defined average height of the dominant and co-dominant trees.

5. **Top height, tree size-ranked** (TH) – refers to the average height of the 100 largest (by DBH) trees per hectare, often used in ground-based inventories. For the data in Table 22, TH=23.95 (m).

6. **Top percent height, tree size-ranked** (TH%) – refers to the average height of the top 5%, 10%, 20% or 30% of the largest DBH trees per hectare or per unit area, often used in ground-based inventories. Other percentages (e.g., 25%, 33%) may also be used. For the data in Table 22:

   TH5 (largest 5%) = 23.95 (m)
   TH10 (largest 10%) = (25.60+23.95)/2 = 24.78 (m)
   TH20 (largest 20%) = (23.42+25.60+23.95)/3 = 24.32 (m)
   TH30 (largest 30%) = (19.25+23.42+25.60+23.95)/4 = 23.06 (m)

   Similar to H%, the average height of the top 5%, 10%, 20% or 30% trees corresponds to the average height of all trees at and above the 95th, 90th, 80th or 70th percentile, respectively, except that H% is ranked by tree height and TH% is ranked by tree DBH.

7. **Lorey's height** (H_L) – refers to the average height of a set of trees weighted by their basal area (BA). For a total of *n* trees, it is calculated as:

   [4.14] $$H_L = \frac{\sum_{i=1}^{n}(BA_i \times HT_i)}{\sum_{i=1}^{n} BA_i} = \frac{\sum_{i=1}^{n}(\pi(DBH_i/200)^2 \times HT_i)}{\sum_{i=1}^{n} \pi(DBH_i/200)^2}.$$

   For the data in Table 22, n=12, $\Sigma BA_i \times HT_i$=11.1082 and ΣBA_i=0.5192. Therefore, H_L=21.39 (m).

   Lorey's height "is often used in remote sensing studies since it provides a measure of forest height that is less affected by thinning and mortality of smaller trees" (Nakai et al. 2010, Erasmi et al. 2019). As shown in [4.14], the calculation of H_L requires the prior knowledge on tree DBH. Tree DBH is typically not a directly classified, observed or extracted variable from any aerial-based remote sensing inventories, i.e., tree DBH or BA must first be predicted indirectly from other models or processes, and H_L must then be calculated from the predictions. As such, it can be highly impacted by other indirect and non-inventory factors. For this reason, some practitioners prefer other stand heights in aerial-based inventories. In any case, readers should exercise caution when using and interpreting Lorey's height, especially when comparing it to other stand heights.

8. **Overstory height** (H_o or H_{DC}) – refers to the average height of the trees equal to or taller than 80% of the tallest tree per unit area. We will discuss this stand height in more detail below.

Other stand heights under various names have also been proposed, such as "crown height", "canopy height", "effective canopy height", "aerodynamic canopy height" (e.g., Nakai et al. 2010), "polygon height", "hexagon height", "mean height" of three tallest or largest trees per plot or per unit area, "predominant height", "dominant height", "site height", and "interpreted height" for a specific set of trees. They can give quite different stand height values for the same stand. So long as they are clearly defined and can be consistently obtained (i.e., repeatable) by different analysts and surveyors on the ground or from an inventory to be compared, there may be values to use them.

For instance, in the above mentioned lidar inventory in Canadian Forest Products Ltd. (Grande Prairie), an area-based hexagon height was obtained from the four tallest trees in each 400 m^2 hexagon, and a polygon height was obtained by averaging the hexagon heights within each polygon (polygon height is stand height in this case). Inherently, the hexagon height and the stand height obtained in this manner are conceptually and methodologically equivalent to the tree height-ranked top height (H_{top}) discussed above.

If necessary, it is always possible to develop predictive relationships between different types of stand heights. An example of such a relationship between tree size-ranked top height and dominant and co-dominant height is available to interested readers. It has been used for many years in Alberta in analyzing data from ground-based inventories.

Calculating Stand Height without Requiring a Full Tree List

For aerial-based remote sensing inventories, because tree DBH is not directly available and because a full tree list may also be very hard to obtain, tree height-ranked top height (H_{top}) is preferred over other heights. One advantage of the H_{top} (the average height of the 100 tallest trees per hectare) is that, it does not require a full tree list, whereas the calculation of the H% requires a full tree list as a priori. Another advantage of the H_{top} is that, an aerial-based inventory technique (e.g., lidar) is much more likely to detect those tall trees, while a tree list from the inventory technique will likely be missing some or many trees, especially small trees.

Recognizing that an aerial-based inventory technique (e.g., lidar) is much more likely to detect tall trees, but usually cannot produce a full tree list accurately, an overstory height based on the "percent to the tallest tree" (i.e., the heights of the trees in percentages relative to the height of the tallest tree) is defined:

- **Overstory height, based on the percent to the tallest tree** (H_o or H_{DC}) – refers to the average height of the trees equal to or taller than 70%, 80% or 90% of the tallest tree per unit area. Other percentages may also be used. For example, for the data in Table 22, the height of the tallest tree is 25.60 (m), and 70%, 80% and 90% of 25.60 are 17.92, 20.48 and 23.04, respectively. Therefore,

  H_o_70 (trees $\geq$ 70% of the tallest tree) = (18.01+18.80+18.96+19.25+19.45+19.65+19.86

  +23.42+23.95+25.60)/10 = 20.70 (m).

  H_o_80 (trees $\geq$ 80% of the tallest tree) = (23.42+23.95 +25.60)/3 = 24.32 (m).

  H_o_90 (trees $\geq$ 90% of the tallest tree) = (23.42+23.95 +25.60)/3 = 24.32 (m).

It is important to note that for instance, the average height of the trees $\geq$ 70% of the tallest tree (H_o_70 =20.70) is entirely different from the average height of the tallest 30% of the trees (H30=23.21). The calculation of the H_o_70 does not require a tree list, whereas the calculation of the H% requires a full tree list (in order to calculate how many trees correspond to the tallest 30% of the trees).

The concept imbedded in the above definition for overstory height can be used to clearly and consistently define the following terms in Table 23 for future studies, where option #1 is preferred and option #2 is sometimes used as a practical simplification, which puts the trees into two classes only (using the distance from the trees to the tallest tree as the criterion). Other options may also be defined (e.g., for multi-layered stand structure) following the same idea, but they are implied in the generic "top A% tallest trees" in option #1. Veterans or residual trees are easy to identify in the early development stages of post-harvest stands. They may be treated as a separate layer or layers. Any distinctly different "super-dominant" trees can be quantitatively defined (e.g., as trees at least 5 m or 130% (whichever is greater) taller than the next tallest tree in a defined layer). They may be considered outliers and treated differently.

TABLE 23. TREE CROWN POSITIONS AND STAND HEIGHT DEFINITIONS FOR AERIAL-BASED FOREST INVENTORIES.

Opt.	Crown position	Tree class (or tree category)	Definition
#1	Overstory (H_o or H_{DC})	Dominant and co-dominant trees	trees $\geq$ 80% of the tallest tree
		Dominant trees	trees $\geq$ 90% of the tallest tree
		Co-dominant trees	80% $\leq$ trees < 90% of the tallest tree
	Intermediate story	Intermediate trees	50% $\leq$ trees < 80% of the tallest tree
	Understory	Suppressed trees	trees < 50% of the tallest tree
	Top A%	Top A% tallest trees	trees $\geq$ (100-A)% of the tallest tree
#2	Overstory	Overstory trees	trees $\leq$ 3 m to the tallest tree
	Understory	Understory trees	trees > 3 m from the tallest tree

Note: opt. denotes option. H_o or H_{DC} is the corresponding overstory height. In option #1, A in top A% denotes a number from 1 to 99.

The trees defined in different classes in Table 23 can be used to calculate the corresponding heights. For instance, for the preferred option #1, the overstory, intermediate story and understory crown positions correspond to the dominant and co-dominant, intermediate and suppressed trees, respectively. In some previous analyses related to this project, the cut-off threshold for separating overstory and intermediate story is 75%. With the refinement to 80%, it is more closely related to the traditional stand height used in Alberta, although other jurisdictions may still choose to use 75% or other reasonable numbers.

The dominant and co-dominant height or overstory height ($H_o = H_{DC}$) calculated according to Table 23 (i.e., from the trees taller than or equal to 80% of the tallest tree) is much more precise, consistent and repeatable than that calculated from the "dominant" and "co-dominant" trees called in the field or interpreted from photos or maps. For this reason, it is highly recommended as a stand height in future studies. One other distinct advantage of the H_o or H_{DC} is that it does not require a full tree list.

When calculating some of the other stand-level heights mentioned earlier (i.e., H_{ave}, H%, TH%, H_L), we typically need a full tree list. For instance, to calculate the tree height-ranked top percent height (H%), the total number of trees must be known and ranked. Otherwise, the average height of the tallest 5%, 10%, 20% or 30% of the trees cannot be calculated. But often, a full tree list may not be available, or is very difficult to obtain accurately, as there may be considerable inaccuracies during the crown delineation or stem segmentation process. The calculation of the overstory or dominant and co-dominant height

according to Table 23 requires the height of the tallest tree. Any other trees whose heights are greater than or equal to the 80% of the tallest tree are counted in. There is no need for a full tree list.

Following the logic embedded in Table 23, other "fraction heights" can also be calculated. For instance, the average height of the trees taller than or equal to the 75, 85% or any other percentage of the tallest tree can be calculated. Fraction heights can be more easily implemented in aerial-based remote sensing inventories than H_{ave}, H% and H_L, as they do not require a full tree list, nor the prediction of tree DBHs.

All stand heights calculated from sample plots are impacted by plot size and no-tally plots. Therefore, plot size and how the plot heights are averaged to derive stand height must be clearly described. Consistence in plot size and stand height calculation is an important consideration when reporting and comparing stand heights. Otherwise, the differences in stand heights can reach several meters or more for the same stand or the same stratum.

To sum up the above discussion, for aerial-based forest inventories, tree height-ranked top height (H_{top}) and dominant and co-dominant height (i.e., overstory height) defined in Table 23 (H_{DC} or H_o) are recommended. Both of these stand heights are precisely defined and thus consistent and repeatable no matter who is using them. Both do not require a full tree list, nor the prediction of tree diameters as a priori. Practitioners can use either one or both of them in practice.

4.5 Stand density in stems per hectare, crown area or crown closure percent

Different stand density measures have been used in different forest inventories. The most common ones are stems per hectare (stems/ha), crown area (m^2) and crown closure/cover percent (CC%).

Data from 28 plots of 400 m^2 each are used to demonstrate the application of the methods described in Sections 4.1 and 4.2. Stand densities in terms of merchantable stems per hectare for these plots (one plot per stand) are obtained through field measurement and from the point cloud segmentation of the lidar inventory in the Canadian Forest Products Ltd. (Grande Prairie). They are listed in Table 24. Since the methods are identical to those demonstrated for tree height, and the only difference is that one continuous variable (tree height) is replaced by another continuous variable (stems/ha), the descriptions below will be brief.

TABLE 24. STAND DENSITIES (STEMS/HA) FROM GROUND MEASUREMENT AND LIDAR INVENTORY.

Plot	Ground	Lidar	Plot	Ground	Lidar	Plot	Ground	Lidar	Plot	Ground	Lidar
1	1100	575	8	175	225	15	1225	675	22	600	775
2	375	400	9	1025	300	16	425	275	23	825	525
3	1200	575	10	300	225	17	475	600	24	375	375
4	350	200	11	900	550	18	550	475	25	150	275
5	450	300	12	975	575	19	400	200	26	450	200
6	625	325	13	600	275	20	475	450	27	1125	525
7	975	550	14	1050	675	21	925	725	28	1475	750

Figure 10 shows the scatter plot of ground density (N_{ground}) against lidar density (N_{lidar}), together with two error plots where the percent error in graph (c) is calculated as $100(N_{ground} - N_{lidar})/N_{ground}$.

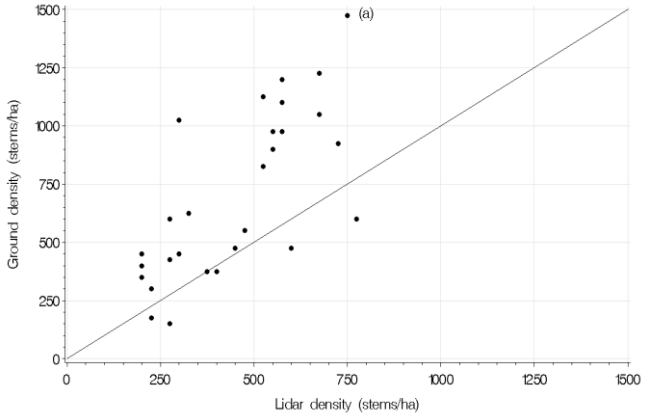

Figure 10 (part 1 of 2). Scatter plot (a) and error plots (b, c) for stand densities measured on the ground (N_{ground}) and from lidar (N_{lidar}). The percent error in graph (c) is calculated as $100(N_{ground} - N_{lidar})/N_{ground}$.

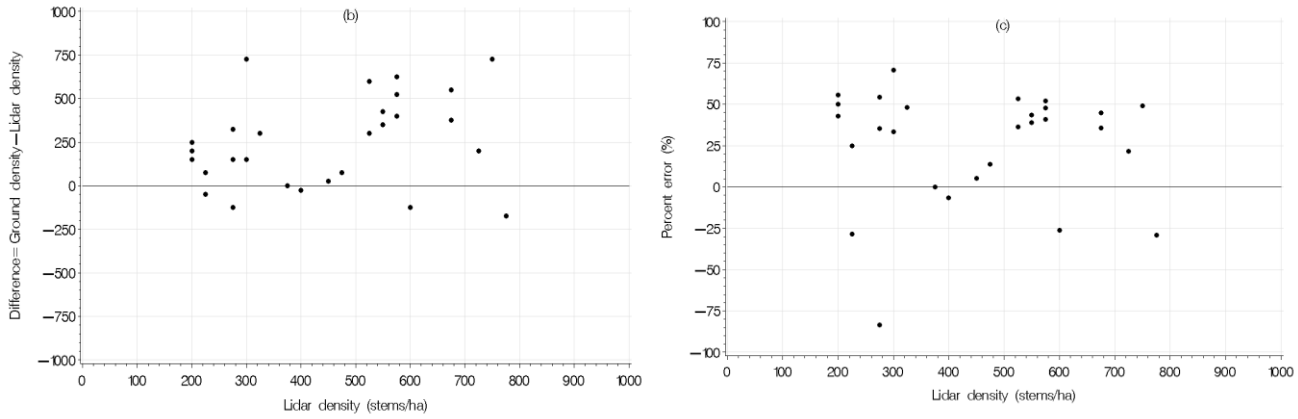

Figure 10 (part 2 of 2). Scatter plot (a) and error plots (b, c) for stand densities measured on the ground (N_{ground}) and from lidar (N_{lidar}). The percent error in graph (c) is calculated as $100(N_{ground} - N_{lidar})/N_{ground}$.

It can be seen from Figure 10 that the differences between ground densities and lidar densities are quite large. Overall lidar densities appear to be underestimating ground densities in most cases (i.e., lidar misses or under-counting stems), but there are cases where lidar densities overestimate ground densities (i.e., lidar over-counting stems).

The large differences between ground and lidar densities are somewhat expected because stem segmentation from lidar point clouds is not an easy task (Ke and Quackenbush 2011). It generally can only find a subset of the merchantable stems on the ground. Smaller and shorter trees are sometimes missed due to their canopy size, spatial position or stand structure. It can be quite challenging to match the GNSS data for a tree collected at the base of a tree to the tree segmented from above through crown detection and delineation, especially when the canopy is quite dense and/or multi-layered.

Table 25 lists the goodness-of-fit statistics and Mielke's measure of agreement for the stand densities. They corroborate and quantify the observations made in Figure 10. Overall, the errors are fairly large (e.g., $\bar{e}$=250.0 stems/ha or $\bar{e}$%=35.8%) and the agreement between ground and lidar densities is quite low (MOA=0.421). Only 11% of the percent errors are within 10% (e_{10}=0.107) and more than 2/3 of the percent errors are greater than 33% (e_{33}=0.321).

TABLE 25. GOODNESS-OF-FIT STATISTICS AND AGREEMENT MEASURE BETWEEN GROUND AND LIDAR DENSITIES.

Type	*n*	$\bar{e}$	MAE	RMSE	$\bar{e}$%	MAE%	RMSE%	e_{10}	e_{33}	e_{50}	MOA
All plots	28	250.0	285.7	356.4	35.8%	40.9%	51.0%	0.107	0.321	0.786	0.421

Note: *n* denotes the sample size (number of plots), $\bar{e}$, MAE, RMSE, $\bar{e}$%, MAE%, RMSE%, e_{10}, e_{33}, e_{50} and MOA are defined in [4.1]-[4.5] and [4.7].

The Bland-Altman plots in actual values and in percentages also show poor agreement between ground and lidar densities (Figure 11). Even though the data points in graph (a) are within the LoAs, there is an obvious upward trend and the absolute values of the errors/differences for many observations are greater than 250 stems/ha. The mean percent error in graph (b) exceeds 25% ($\overline{PE}$%= 25.9%), and six out of the 28 percent errors are outside $\pm$50%.

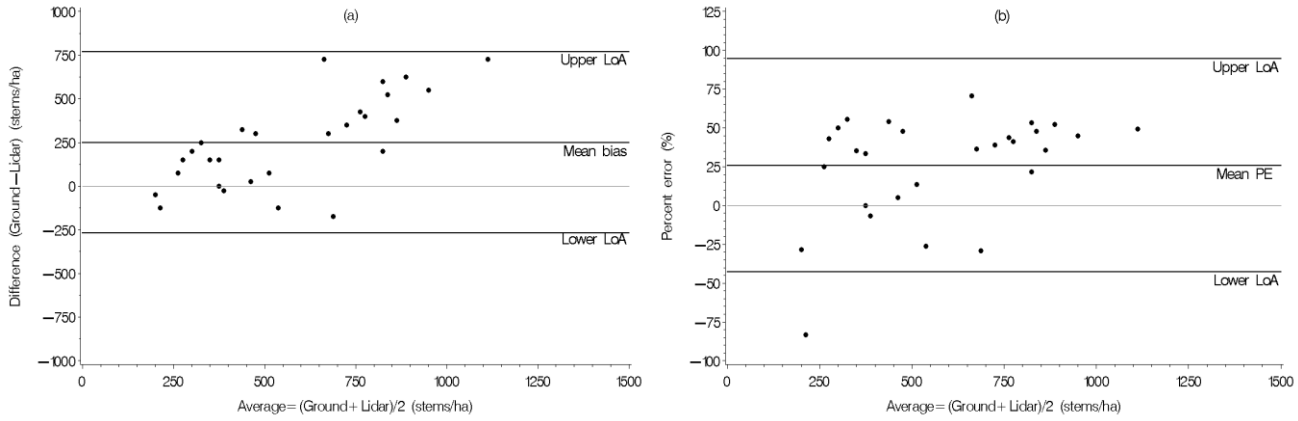

Figure 11. The Bland-Altman plots in actual values (a) and in percentages (b) for stand densities from the ground and lidar measures. Actual data are listed in Table 24.

The cumulative distributions for the ground densities and lidar densities clearly show the divergence between the two sets of densities (Figure 12).

The KS test between the ground and lidar densities shows that the maximum distance between the two cumulative distributions is D=0.4286, which occurs at the density of 775 stems/ha. With the known D, the z is calculated to be: $z=0.4286\sqrt{28\times 28/56}$=1.604 (following [4.12]). Therefore, the p-value=0.0117 (following [4.11] or [4.13]). This p-value is smaller than α=0.05, indicating that the frequency distribution of the lidar densities is different from that of the ground densities.

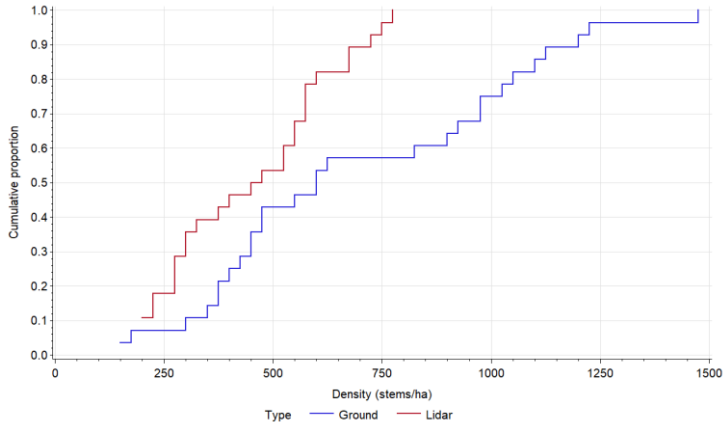

Figure 12. Cumulative distributions for stand densities from the ground and lidar measures. Actual data are listed in Table 24.

Based on all of the above assessments (from Figures 10-11, Table 25 and the KS test), it can be inferred that for this example, the agreement between ground densities and lidar densities is not good. Lidar densities appear to substantially underestimate ground densities in most cases. Without any adjustment or calibration, lidar densities are not representative of the ground densities. They should not be used to substitute the ground densities unless they are adjusted or calibrated.

Stand Density in Terms of Crown Area or Crown Closure Percent

Instead of using stems/ha, if stand density is expressed in terms of crown area (e.g., 160.98 m^2/ha), the analysis follows the standard analysis demonstrated above for continuous variables.

If stand density is expressed in terms of (exact) crown closure percent (CC%, i.e., % of ground covered if crown projected vertically from above, e.g., 6%, 80%, 92%), the analysis still follows the standard analysis for continuous variables, except that extra care must be exercised when calculating and interpreting the average crown closure percent from different crown closure percents, which are ratios or proportions that do not reveal their numerators nor their denominators. Without using a consistent basis or without knowing the abundances in terms of the actual numbers (numerators and denominators), averaging the CC% is not meaningful.

If, for whatever reason, stand density is expressed in terms of crown closure classes defined by the ranges of crown closure percent (such as those A, B, C and D density classes in the Alberta Vegetation Inventory), stand density becomes a categorical variable. As such, the analysis appropriate for categorical variables should be implemented.

4.6 Goodness-of-fit measures for categorical variables

In theory, the goodness-of-fit statistics, the *agreement measure and the plots described in Section 4.1 could be applied to* categorical variables as well. For instance, Table 26 lists the calculated goodness-of-fit statistics and Mielke's measure of agreement based on the data in Table 3, comparing the ground counts (y=[55, 18, 0, 8, 14, 10, 9, 28, 4, 63]) to the corresponding classification counts from the lidar inventory (x=[57, 19, 2, 14, 19, 10, 9, 25, 2, 52]). Scatter plot, error plots and the Bland-Altman plot could also be drawn from these counts (available to interested readers).

TABLE 26. GOODNESS-OF-FIT STATISTICS AND AGREEMENT MEASURE FOR GROUND VS CLASSIFICATION COUNTS.

Type	*n*	$\bar{e}$	MAE	RMSE	$\bar{e}$%	MAE%	RMSE%	e_{10}	e_{33}	e_{50}	MOA
Ground vs classification	10	0	3.20	4.52	0	15.3	21.6	0.444	0.667	0.889	0.976

Note: n is the sample size and all statistics are defined in [4.1]-[4.7]. Actual data from the ground (y) and classification (x) are listed in Table 3.

However, since the categories of a categorical variable often do not exceed 10 in our studies (e.g., the number of species or the number of crown closure classes typically does not exceed 10), the statistics calculated from small sample sizes may not be reliable. For this reason, we generally do not recommend the calculation of the goodness-of-fit statistics and *agreement measure for* categorical variables. The statistics and the methods described in Sections 2 and 3, and the logic embedded therein, specifically the error matrix and the chi-square test or Fisher's exact test, are designed to better fit categorical variables. They are more appropriate and should be sufficient in determining the accuracy and comparing the agreement (to ground measures) for categorical variables. Interested readers may wish to read Agresti (2013, 2018) for some more complex treatments of analyzing categorical variables.

4.7 Goodness-of-fit measures for ground = *f* (inventory variables) models

The methods and the goodness-of-fit statistics, the agreement measure and the plots described in previous Sections apply to *relevant* continuous and categorical variables obtained from any inventory approach, be it tree-based, area-based or a hybrid of the two.

Often, photogrammetric and remote sensing studies involve the development of regression relationships (or models) between ground measures and inventory measures (Næsset 2002; Næsset et al. 2005, 2013; Vastaranta et al. 2012; White et al. 2017; Puliti et al. 2017; Surový and Kuželka 2019; Coops et al. 2021), where ground measures refer to ground-measured forest attributes and inventory measures refer to inventory-derived or extracted measures, variables or metrics (e.g., lidar-derived measures or variables we typically use as predictors are based on lidar-derived height, crown cover and vertical structure metrics). This is particularly true for area-based approach, where the inventory technique is almost entirely dependent on the development of regression models linking inventory-derived variables or metrics (*x* variables or predictors) to ground-measured forest attributes such as height, density, diameter, volume and biomass (*y* variables), based on the following general form:

[4.15] Ground = *f* (inventory variables)

where ground denotes the ground measurements for forest attributes, *f* denotes a linear or nonlinear function and inventory variables denote the inventory variables, measures or metrics (e.g., lidar metrics) extracted or derived from an inventory technique.

For example, at the individual tree level:

[4.16] $HT_{ground} = f\ (HT_{lidar})$

[4.17] DBH = *f* (lidar height, crown cover and vertical structure metrics)

where HT_{ground} is the tree height measured on the ground, HT_{lidar} is the corresponding lidar-derived height estimate, often called lidar height, and DBH is the tree diameter (at breast height) measured on the ground. If warranted, other lidar metrics may also be included in [4.16], but HT_{lidar} is usually enough and we will just use [4.16] to illustrate the general concept and logic here. Tree taper and different types of tree volumes can be obtained once HT_{ground} and DBH are known (Huang 1994). They can also be predicted directly from lidar metrics or lidar-derived variables via Tree taper = *f* (lidar metrics or variables).

At the stand level (i.e., for area-based approach):

[4.18] Stand height = *f* (lidar height variables)

[4.19] Volume (m^3/ha) = *f* (lidar height, crown cover and vertical structure variables)

[4.20] Other stand level variables = *f* (lidar metrics or variables)

where stand height can be H_{DC}, H_{top} or other stand heights discussed in Section 4.4, and other stand level variables in [4.20] can be stand density (stems/ha), basal area (m^2/ha), average DBH (cm), height and diameter distributions, aboveground biomass (Mg/ha, Mg = megagram =1,000 kilograms), CC%, age, site index, etc. Once again, if warranted, other lidar metrics or variables may also be included in [4.18].

For simplicity, only continuous *y*-variables are illustrated in [4.16]-[4.20]. Dichotomous (binary) variables with only two possible outcomes, such as "correct/incorrect" for species, "live/dead" for mortality and "yes/no" for ingrowth, are not discussed here. Species prediction typically involves machine learning and a large number of inventory technique-derived metrics (e.g., Forsite Consultants Ltd. 2020, Hologa et al. 2021). Mortality and ingrowth predictions require some specialized regression techniques (e.g., Yang et al. 2003; Yang and Huang 2013, 2015; Cortini et al. 2017). Interested readers can find more details in these references.

When developing regression models expressed in [4.15]-[4.20], two of the commonly seen goodness-of-fit statistics in photogrammetric and remote sensing studies are regression root mean square error ($RMSE_r$) or mean squared error (MSE_r), and coefficient of determination (R^2):

[4.21] $$\text{RMSE}_r = \sqrt{\frac{1}{n-p}\sum(y_i - \hat{y}_i)^2} = \sqrt{\frac{\sum e_i^2}{n-p}} \quad (\text{or } \text{MSE}_r = \frac{\sum(y_i-\hat{y}_i)^2}{n-p} = \frac{\sum e_i^2}{n-p})$$

[4.22] $$R^2 = 1 - \frac{\sum(y_i-\hat{y}_i)^2}{\sum(y_i-\bar{y})^2}$$

where $\hat{y}_i$ is the predicted y_i from the fitted regression model $\hat{y}_i = f$ (x variables) (or Ground = f (inventory variables)), $e_i = y_i - \hat{y}_i$ is the residual, n is the number of observations, p is the number of parameters related to the x-variables, and (n - p) is the error degrees of freedom from the model with p parameters.

Notice the important difference between the $RMSE_r$ in [4.21] (which involves the regression model $\hat{y}_i = f$ (x variables) and needs to be corrected by the error degrees of freedom from the fit) and the RMSE in [4.3] (which is directly calculated from the data). Since $RMSE_r$ can be highly influenced by the number of parameters (or variables) used in the developed regression model (and sometimes, a large number of variables may be used as x variables in lidar predictions), the $RMSE_r$ reported in many studies may not be as meaningful as the RMSE in [4.3] in assessing inventory techniques.

Notice also the difference between the $RMSE_r$ and the unbiased estimate of the standard deviation of residuals (SD_r) in regression analysis (where the mean of the residuals is $\bar{e} = \Sigma(y_i - \hat{y}_i)/n = 0$):

[4.23] $$\text{SD}_r = \sqrt{\frac{1}{n-1}\sum(e_i - \bar{e})^2} = \sqrt{\frac{\sum e_i^2}{n-1}} = \sqrt{\frac{1}{n-1}\sum(y_i - \hat{y}_i)^2}\ .$$

Interested readers may want to read Huang et al. (2019) to see the exact mathematical relationships and conversions among $RMSE_r$, SD_r and RMSE (and SD when $\bar{e} \neq 0$). They should be understood correctly and used in the right contexts clearly and consistently, especially when comparisons are made or when SD_r is referred to as "RMSE" or the unbiased estimate of RMSE.

Numerous studies also used R^2 (or Pearson's correlation coefficient, r) to show how accurate and reliable the regression models and the inventory technique were. This was likely caused by the misunderstanding of the R^2 (or r), which only describes the correlation between the variables (y and x), but not the accuracy and agreement (to ground measures) of an inventory technique. In fact, an inaccurate or a completely wrong or mismatched inventory technique relative to ground measures can produce a very high R^2 value when Ground = f (inventory variables) is developed. This can be illustrated using Figure 13, where three hypothetical data sets of ground (y) versus inventory (x) measures for a variable (e.g., height) from three inventory techniques (T1, T2 and T3) are shown.

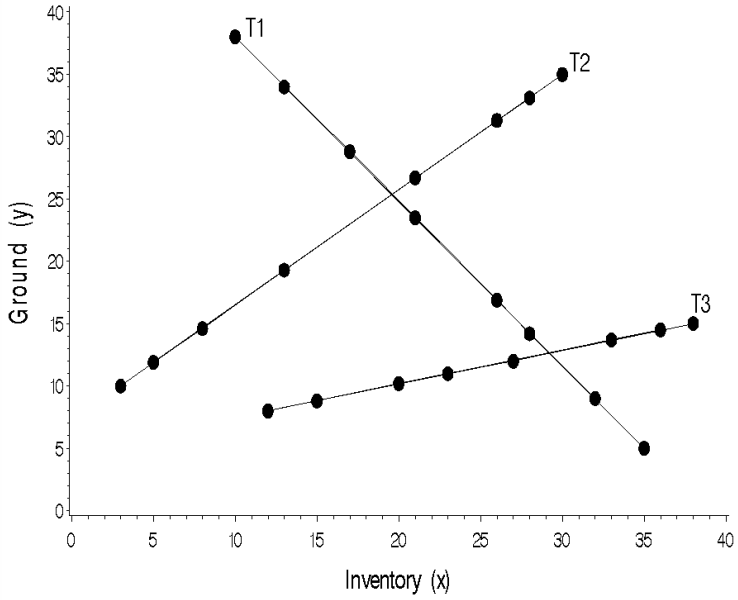

Figure 13. An illustration of three data sets between ground (y) and inventory (x) measures from three inventory techniques (T1, T2 and T3), each with an R^2 value of 1, but the accuracies and the agreement (to ground measures) of the three inventory techniques are completely different.

Since the data lie on the straight lines in Figure 13, each differing data set has an R^2 (or r) value of 1. However, it is without question that the three data sets from three inventory techniques correspond to very different values and display entirely different agreement patterns between ground measure (y) and inventory measure (x), even though all of them have an R^2 value of 1.

It is critically important to recognize (and reiterate) that for any developed regression model from any inventory approach, regardless of how high the R^2 value is, it only measures the correlation between y and x, not the accuracy, nor the agreement (or "closeness", "likeness", "concordance", "equivalence", "same-ness") between ground measures and inventory variables from an inventory technique. Therefore, the R^2 must be interpreted with extreme caution, if it is used at all.

To elaborate the above statement further using tree height measurements as an example, assume that the tree height measurements from lidar were consistently twice as high as the tree height measurements on the ground. If we were to plot the tree heights on the ground (y) against the tree heights from the lidar (x) in the style of Figure 13, we would get a perfect straight line with an R^2 of 1 and a slope of 0.5. If the tree heights from the lidar were consistently half of what the tree heights on the ground are, we would still get a perfect straight line with an R^2 of 1 but a slope of 2.0. The R^2 in both cases would be 1 even though in one case the tree heights from the lidar were twice as high as the tree heights on the ground, and in the other case they were only half of what the tree heights on the ground are. Clearly the tree heights from the lidar did not match or agree with those on the ground in both cases but the R^2 values would suggest that the "agreement" between lidar heights and ground heights were perfect.

Indeed, a high R^2 value for the tree height example only implies that the height measurements between ground and lidar correlate well with each other. But it does not mean that the measurements agree with each other, or the lidar heights are accurate and reliable (although one could argue that for this example lidar heights were reliably inaccurate and wrong). The R^2 only measures the correlation but not the agreement between the two measurements, nor the accuracy and validity of an inventory technique. We will have a perfect correlation (R^2=1) as long as the data points lie along any straight line, in any direction without anything to do with how good or bad the lidar heights actually match the ground heights. Entirely opposite or completely wrong or reversed lidar heights to ground heights can still produce a very height R^2 value. The same argument applies to any other variables and models expressed in [4.15]-[4.20].

The R^2 is very useful in judging regression correlation, but is not really relevant (and is considered flawed by many researchers) in judging the agreement between any two sets of values linked through regression analysis. Many of the reported high R^2 values in remote sensing and photogrammetric studies (e.g., Næsset 1997, Means et al. 1999, Lim et al. 2003, Heurich et al. 2004, Næsset et al. 2005, Kwak et al. 2007, Sibona et al. 2017, Wang et al. 2019) are only relevant to the correlation, but not to the agreement between the two sets of measurements even though sometimes it is interpreted to be so and used to showcase the accuracy and validity of an inventory technique.

Methodologically and fundamentally, it is inappropriate to use the R^2 value to judge the agreement between any two sets of measurements. We cannot say, for instance, that the lidar heights match or agree with the ground heights even if the R^2 value between ground heights and lidar heights is 1.

We need to use the methods, the goodness-of-fit statistics, the agreement measure and the plots described in previous Sections, not the $RMSE_r$ (or MSE_r) and R^2 from regression analysis, to evaluate the accuracy and agreement (to ground measures) of an inventory technique. The statistics from regression analysis are more suited to method or model calibration in the development of inventories, or for corrections when inventory variables/measures do not agree with ground measures *(see Section 5.6).* Practitioners must recognize that judging the accuracy and agreement of an inventory technique relative to ground measures (i.e., agreement analysis), is very different from judging the regression models between ground measures and inventory measures (i.e., correlation analysis). They should not be confused even though they have been in many research papers. Agreement and correlation are two entirely different concepts (Robinson 1957; Altman and Bland 1983, 1987; Bland and Altman 1986, Hollis 1996; Stehman 1997; Liao and Lewis 2000, Ludbrook 2002, Bunce 2009, Choudhary and Nagaraja 2017).

Inherently the futility of R^2 in judging the agreement between ground and inventory measures is due to the invariance property of the R^2 from changes in location, direction as well as scale in the data. In plain languages the invariance property of the R^2 means that, if we shift the data up or down, left or right, or if we tilt the data by multiplying/dividing and adding/subtracting some *arbitrary* or *random* numbers to the original data to form some new variables and completely different agreement patterns and relationships between ground measures and inventory variables (e.g., $y_1 = 10.88y - 3.65$ and $x_1 = x/2 + 138$, or $y_2 = y/13 + 80.48$ and $x_2 = 12x - 100.28$, for any given y and x values – such as those from Data-1 in Table 16), the R^2 from the ordinary least squares (OLS) fits of $\hat{y}_1 = a_1 + b_1x_1$ and $\hat{y}_2 = a_2 + b_2x_2$ will be the same, and both will be identical to that from $\hat{y} = a + bx$ (interested readers should test this out, the R^2 in this case for Data-1 in Table 16 always equals to 0.9137, even

though the *y*-*x* data are arbitrarily changed to y_1-x_1 or y_2-x_2). The implications of this very important but often overlooked property are twofold:

1. The R^2 cannot adequately reflect the changes in agreement patterns between ground measures and inventory variables/measures from different inventory techniques; and

2. Even though ground measures and inventory variables differ, disagree or mismatch greatly, the R^2 from Ground = *f* (inventory variables) can remain the same and appear very good (e.g., close to 1), which can be mistaken as good agreement between ground measures and inventory variables from an inventory technique.

Two other goodness-of-fit statistics frequently seen and called “agreement measures” in remote sensing and photogrammetric studies, the index of agreement (Willmott 1981, 1982) and the agreement coefficient (Ji and Gallo 2006), also possess some critical drawbacks. They may not be as easy to comprehend as those for the R^2. We provide some additional notes for interested readers in Section 5.3.

4.8 A note of caution on testing intercept and slope in scatter plot

One of the most common graphical means for assessing the accuracy and agreement of any variables is the plotting of the scatter plot of *y* against *x* (e.g., Figure 3(a), Figure 7(a) and Figure 10(a)), where *y* denotes the “truth” or ground reference and *x* denotes the prediction or estimation from an inventory. A widely held belief in accuracy assessment and agreement studies is that, if the predicted values agree with the true values, a simple linear regression between the two sets of values should follow the 45° line that passes through the origin (i.e., *y*=*x*). Therefore, in order to ascertain that the data in the scatter plot follow the 45° line, it is very tempting to fit a simple linear regression to the data, then evaluate if the intercept is 0 and the slope is 1 for the fitted line.

For instance, following the standard ordinary least squares (OLS) method (e.g., Draper and Smith 1998), a simple linear regression can be fitted for the data in Figure 3(a):

[4.24] $\hat{y}_i = a + bx_i$

where the estimated intercept *a*=-1.13343 and the slope *b*=1.09234. Figure 14(a) shows the original data (listed as Data-1 in Table 16) and the fitted regression line.

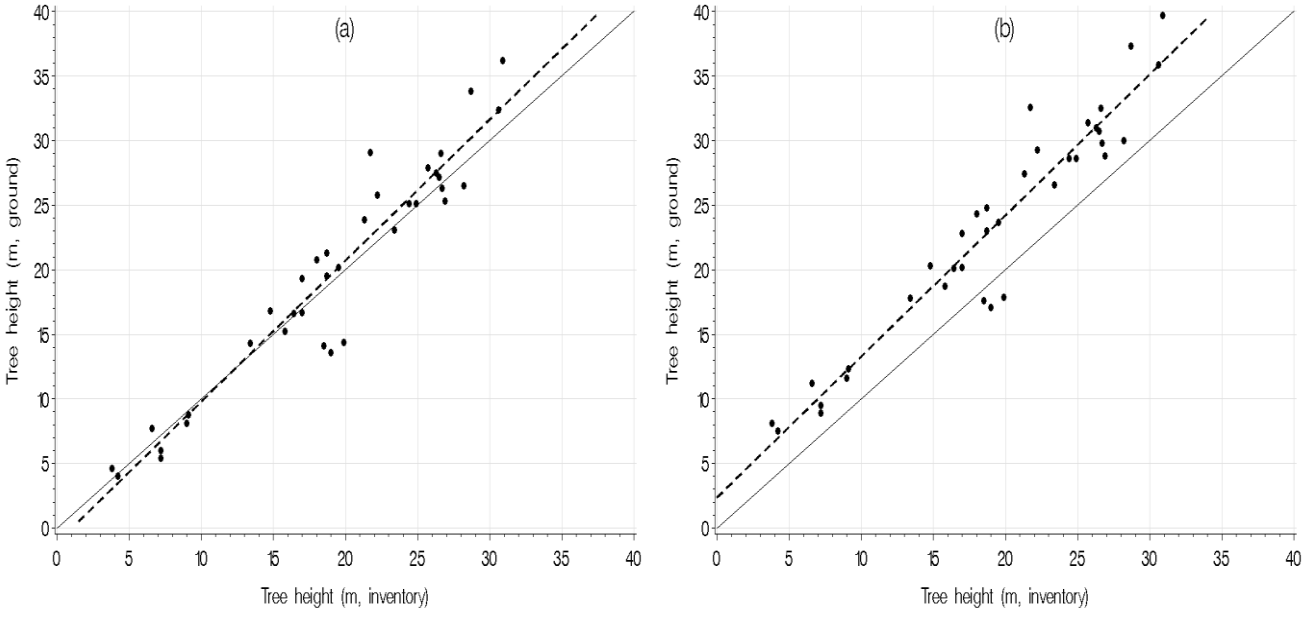

Figure 14. Simple linear regressions (dashed lines) between ground (*y*) and inventory (*x*) tree height data. In graph (a), Data-1 in Table 16 are used. In graph (b), the same Data-1 are used, but with 3.5 m added to every *y* value (the *x* values remain the same). The solid lines are the 45° lines. Testing the fitted regressions ($\hat{y}$=*a*+*bx*) would lead to the acceptance of *a*=0 and *b*=1 in both cases.

Testing whether *a*=0 and *b*=1 for the fitted line in Figure 14(a) can be done through a simultaneous *F*-test, separate *t*-tests, or confidence interval inferences. Details on the tests and inferences were provided elsewhere (Huang 2002, Huang et al. 2019). They will not be repeated here. Results showed that we cannot reject *a*=0 and *b*=1 (at α=0.05).

Interestingly though, for instance, when 3.5 m is added to every *y* value in the data to form a new variable y_1 (i.e., $y_1=y+3.5$ but the *x* remains the same, as in Figure 14(b)), re-fitting the simple linear regression $\hat{y}_1=a+bx$ will produce an intercept *a*=2.36657 and a slope *b*=1.09234. Testing whether *a*=0 and *b*=1 for the new regression would still show that we cannot reject *a*=0 and *b*=1 (Huang et al. 2019), even through the differences between the data sets and the regressions in Figure 14(a) and Figure 14(b) are clearly different.

Intuitively, it seems logical that in accuracy assessment and agreement analysis, if the predicted (or indirectly measured) values agree with the true values, a simple linear regression expressed as [4.24] between the two sets of values in the scatter plot should be a 45° line through the origin. Therefore, it seems natural to test if *a*=0 and *b*=1 for the fitted regression. But as early as 1972, Aigner (1972) pointed out that such an intuition is generally wrong when comparing true values to indirectly predicted or measured values from a device, a technique or a model. Harrison (1990), Kleijnen et al. (1998) and Kleijnen (1999) provided some other explanations and technical details as to why such a test is wrong. In fact, Kleijnen et al. (1998) and Kleijnen (1999) called it the "naïve test" and listed it as an example of a wrong approach.

While such labelling might be hard to accept, especially since we have used it many times in some earlier studies (e.g., Huang et al. 1999), a more reasoned assessment of the testing does suggest that such testing is not appropriate in assessing the accuracy and measuring the agreement between *y* and *x*. While this conclusion may appear counter-intuitive with regard to the scatter plot, it is the right one.

The danger of testing whether *a*=0 and *b*=1 lies more on the fact that, it frequently leads to wrong conclusions by accepting invalid models and rejecting valid models. Accepting invalid models is relatively easy to see (e.g., Figure 14(b)). To understand the latter (which requires some additional statistical and mathematical formulations that some practitioners may not be familiar with, but we provide here for interested readers), recognizing that the parameters (*a* and *b*) in [4.24] are computed as follows (Neter et al. 1989, pp.102-103; Draper and Smith 1998, p.42; Huang et al. 2019):

[4.25] $$b = r\sqrt{\frac{\sum(y_i-\bar{y})^2}{\sum(x_i-\bar{x})^2}} = r\sqrt{\frac{\sum(y_i-\bar{y})^2/n}{\sum(x_i-\bar{x})^2/n}} = r\sqrt{\frac{S_y^2}{S_x^2}} = r\frac{s_y}{s_x}$$

[4.26] $$a = \bar{y} - b\bar{x}$$

where $S_y^2 = \sum(y_i-\bar{y})^2/n$ and $S_x^2 = \sum(x_i-\bar{x})^2/n$ are the variances and S_y and S_x are the standard deviations for the *y* and *x* values, respectively; $\bar{y}$ and $\bar{x}$ are the averages of the *y* and *x* values, respectively; and *r* is Pearson's correlation coefficient (or simply correlation coefficient) calculated by:

[4.27] $$r = \frac{\sum(y_i-\bar{y})(x_i-\bar{x})}{\sqrt{\sum(y_i-\bar{y})^2\sum(x_i-\bar{x})^2}}$$

If the models are excellent in the sense that the ground values and the inventory values have the same positive mean $\mu_y=\mu_x$ ($=\mu>0$) and the same variance $S_y^2=S_x^2$ ($=\sigma^2$), they can still lead to $a\neq0$ and $b\neq1$ in [4.24], because in reality, any inventory will not give perfect estimates (i.e., matching the true values exactly or $y=x$) for all values, no matter how good the estimates are from the inventory. Hence, $0 < r < 1$, which means:

[4.28] $$0 < b = r\frac{s_y}{s_x} = r\frac{\sigma}{\sigma} < 1$$

[4.29] $$0 < a = \bar{y} - b\bar{x} = \mu - b\mu = \mu(1-b) < \mu$$

Therefore, if a test of *a*=0 and *b*=1 is conducted, it will likely reject *a*=0 and *b*=1.

As mentioned before (in Section 4.1), the scatter plot is a good starting point in assessing the accuracy and agreement of attribute estimates from an inventory technique. It provides a first impression and helps the eye in gauging the degree of accuracy and agreement between the ground truth and inventory estimates. However, it has several limitations and is not sufficient in judging the accuracy and agreement of an inventory technique or any models derived from the inventory technique. Focusing on a test or a method that proves the data in the scatter plot have an intercept of 0 and a slope of 1 is certainly not the way to go. Readers who are interested in more technical details about this and several related topics may wish to read Kleijnen et al. (1998), Huang (2002), Yang et al. (2004), Piñeiro et al. (2008) and Huang et al. (2019).

5 Additional notes

The following additional notes are provided for interested readers who may want to know the reasons and justifications for many of the concepts and methods described in this study. They also provide some added technical details and caveats on several more intricate and easily entangled concepts and methods. In addition, they describe the general methods for calibration, which may be necessary when the predictions from an inventory technique are poor and when adjustments and corrections are needed.

5.1 Clarification on confusion matrix

In many conventional remote sensing studies, a table similar to Table 2 or Table 3 has often been called "confusion matrix" or "confusion table" (Story and Congalton 1986; Lillesand et al. 2015; Foody 2002, 2020). It is very common to use such a table to represent the classification accuracies of remotely sensed data and as the basis for further analysis. There may not be anything fundamentally wrong in calling such a table "confusion matrix" and use it to derive the "omission error", "producer's accuracy", "commission error" and "user's accuracy" or "reliability", except that they can be confusing even to researchers and specialists publishing in peer-reviewed journals (e.g., Scofield et al. 2015, Radoux and Bogaert 2017), let alone many non-academic practitioners.

To illustrate, a simple data set of two plots is used. It is listed in Table 27, where "ground" refers to the ground observations and "inventory" refers to the corresponding classifications from an inventory technique. Ground observations are considered to be the truth and used as the reference standards.

TABLE 27. GROUND AND INVENTORY DATA FROM TWO EXAMPLE PLOTS (FOR ILLUSTRATION).

Plot	Tree	Ground	Inventory	In/correct	Plot	Tree	Ground	Inventory	In/correct
1	1	Sw	Sw	Correct	2	1	Sw	Sw	Correct
1	2	Sw	Sw	Correct	2	2	Sw	Sw	Correct
1	3	Sw	Sw	Correct	2	3	Sw	Fb	Incorrect
1	4	Sw	Sw	Correct	2	4	Aw	Pl	Incorrect
1	5	Sw	Sb	Incorrect	2	5	Aw	Aw	Correct
1	6	Sw	Fb	Incorrect	2	6	Aw	Pb	Incorrect
1	7	Sb	Fb	Incorrect	2	7	Aw	Aw	Correct
1	8	Sb	Fb	Incorrect	2	8	Aw	Pb	Incorrect
1	9	Sb	Fb	Incorrect					
1	10	Sb	Pl	Incorrect					

Note: ground denotes ground-observed species, inventory denotes inventory-classified species, species are as defined in Table 1, and in/correct denotes correct or incorrect classification by the inventory.

Based on the data in Table 27, a "confusion matrix" is formulated in Table 28. We use the conventional terminologies for now, and list the "producer's accuracy" and "user's accuracy or reliability" in Table 28.

TABLE 28. CONFUSION MATRIX CORRESPONDS TO THE TWO EXAMPLE PLOTS IN TABLE 27.

		Inventory						Total	Producer's
	Species	Sw	Sb	Aw	Pb	Fb	Pl	**(row)**	**accuracy**
Ground (reference)	Sw	6	1	0	0	2	0	9	67%
	Sb	0	0	0	0	3	1	4	0
	Aw	0	0	2	2	0	1	5	40%
	Pb	0	0	0	0	0	0	0	N/A
	Fb	0	0	0	0	0	0	0	N/A
	Pl	0	0	0	0	0	0	0	N/A
Total (column)		6	1	2	2	5	2	18	
User's accuracy (reliability)		100%	0	100%	0	0	0		

Note: species are defined in Table 1, and N/A denotes not available or not applicable (i.e., due to a denominator of zero). The overall accuracy for all species combined is P_o = 8/18 = 44%.

The "omission error" and "commission error" are complementary to the producer's accuracy and user's accuracy, respectively (i.e., producer's accuracy + omission error = 100%, user's accuracy + commission error = 100%). They are not listed in Table 28.

Having a “confusion matrix” like Table 28, many confusions could occur. Story and Congalton (1986) and Congalton (1991) provided good examples in which the producer’s accuracy might be misinterpreted. Here we explain the confusions that could occur from a different angle, using the user’s accuracy or the so-called reliability. First though, the concepts of omission error and commission error need to be clarified.

Omission error (or error of omission) – generally refers to the ground samples that are left out or “omitted” from the classification. It occurs when the ground samples of a species are mistaken to be another species by the classification. Omission error is derived based on the values listed in the confusion matrix. Species or stems missed/omitted by crown deliberation/stem segmentation are not counted as omission error (see Sections 2.2 and 2.3). Omission error for each species is calculated by adding together the incorrect classifications for the species relative to the ground observations and dividing them by the total number of ground observations for that species. For instance, for Sw in Table 28, the omission error is (1+2)/9=33%. Omission error for each species is complementary to the corresponding producer’s accuracy. Hence, it can also be calculated as: omission error = 100% - producer's accuracy = 100% - PR.

Commission error (or error of commission) – an omission error for one species becomes a commission error for another species. Commission error for each classified species is calculated by adding together the incorrect classifications for the species relative to the classifications and dividing them by the total number of classifications for the species. For instance, for Sw in Table 28, the commission errors are 0/6 = 0%. Commission error for each species is complementary to the corresponding user's accuracy. It can also be calculated as: commission error = 100% - user's accuracy = 100% - PC.

From Table 28, the overall accuracy is calculated to be (6+0+2+0+0+0)/18=44%. This level of accuracy is generally considered low. However, for instance, if a user is most interested in white spruce (Sw), the user’s accuracy or reliability for Sw is calculated to be 100% (6/6), which is perfect! This might lead the unsuspecting user to conclude that, although this classification has an overall accuracy that is considered low, it is perfect and 100% reliable for Sw. Making such a conclusion could be a serious mistake. A quick calculation of the producer’s accuracy for Sw gives a value of 67% (6/9), which is much less than perfect. In other words, although 100% of the Sw on the map have been correctly identified as Sw on the ground, and a user of this map may claim a “100% reliability” for Sw when he/she uses this map in the field, only 67% of all Sw on the ground actually appeared on the map as Sw. The other 33% (omission error) are mistaken to be other species. Clearly, a “100% reliability” or a “perfect user’s accuracy” may not mean much. It must be weighed and interpreted with great caution depending on the use of the map.

Similarly, for Aw in Table 28, the user’s accuracy is also 100% (2/2). This might lead an unsuspecting user to conclude that the classification is perfect and 100% reliable for Aw. Making such a conclusion could again be a serious mistake. The producer’s accuracy for Aw is only 40% (2/5), which is much less than perfect. Therefore, although a user of this map can claim that 100% of the Aw appeared on the map have been correctly identified as Aw on the ground, only 40% of all Aw on the ground actually appeared on the map as Aw. The other 60% are misclassified as other species on the map.

The risks of emphasizing the conventional “user’s accuracy” and calling it “reliability” in some literature can be more profound to many unsuspecting users or to any users, as a “100% reliability” or a “100% user’s accuracy” could be easily (and mistakenly) thought to be 100% accurate for the classification. This is highlighted further in a simpler example with three species only (Table 29).

In Table 29, the “user’s accuracies” for both Sw and Sb are 100%. It would be a complete travesty if they are used to claim “100% perfect reliabilities” for these two species without recognizing that only 3 out of 18 (17%) white spruce and 8 out of 15 (53%) black spruce are correctly classified and actually appeared on the map. Unfortunately, it is not uncommon to see that the classifications were poor, yet the calculated “user’s accuracies” or “reliabilities” were very high.

TABLE 29. AN ILLUSTRATION OF A CONFUSION MATRIX AND SOME RELATED CALCULATIONS.

	Species	Inventory			Total (row)	Producer’s accuracy
		Sw	Sb	Fb		
Ground (reference)	Sw	3	0	15	18	17%
	Sb	0	8	7	15	53%
	Fb	0	0	40	40	100%
Total (column)		3	8	62	73	
User’s accuracy		100%	100%	65%		

For Fb in Table 29, a 100% “producer’s accuracy” just means that, on paper (map), 100% of the Fb have been correctly identified as Fb, but a user of this map will find that only 65% (40/62) of the time that the map says is Fb will actually be Fb on the ground.

Some readers may already have noticed that, while we spent considerable length in this section on “confusion matrix” and “confusion matrix” related terminologies like “producer’s accuracy” and “user’s accuracy” or “reliability”, we only briefly mentioned them in the main text (in Section 2.1). We purposely avoided using these conventional idioms in the main text for the following reasons:

1. A term like “error matrix” or “classification performance matrix” is more intuitive to most practitioners, and more pertinent and informative to what we try to convey than the “confusion matrix”. Prior to being called “confusion matrix” in remote sensing studies, the earliest example we found for a tabulation of classification results is called “error matrix” (Aronoff 1982). There is really no need to convolute it and call it “confusion matrix” to redefine what Aronoff (1982) and other researchers have already clearly defined.

2. The term “confusion matrix” or “table of confusion” originates from machine learning and computer science (https://en.wikipedia.org/wiki/Confusion_matrix). It typically only deals with two categories or two classes in such a table (i.e., 2×2 contingence tables or matrices), to see whether a classification is confusing two classes (i.e. mislabeling one as another). However, a typical confusion matrix in machine learning involves many theoretical concepts that we try to avoid in this study, such as condition positive, condition negative, true positive, true negative, false positive/type-I error and false negative/type-II error. Numerous accuracy and error measures can be derived from such a confusion matrix, such as “threat score”, “markedness”, “prevalence threshold”, “F1 score”, “harmonic mean of precision and sensitivity”, “balanced accuracy”, and “Matthews correlation coefficient”. They may have their roles to play in machine learning, but they are irrelevant in indicating the classification accuracy of any inventory technique. We will briefly discuss many of them later (Section 5.2).

3. "Producer’s accuracy” was used to measure the omission error. "User’s accuracy” or “reliability” was used to measure the commission error. As illustrated in Tables 28-29, both “producer’s accuracy” and “user’s accuracy” could be very misleading and easily confused or misunderstood. There is really no need to call, for example, “what percent of the white spruce trees in the ground sample were correctly classified as white spruce in the inventory”, as “producer’s or user’s accuracy. It could be either one. In spite of the fact that researchers wrote research papers to clarify the difference between "producer’s accuracy” and "user’s accuracy”, researchers still confused and mixed/reversed between the two terms (see Section 5.2). Of course, it could be argued that it was the persons who confused the terms. But it really does not make sense to ask non-academic practitioners to combine these two easily confused or misunderstood terms to come up with something that is clear and less confusing in practice. It would be very hard for practitioners to make two negatives a positive.

4. The original concepts of “producer's risk” and “consumer's risk” from a branch of statistics known as acceptance sampling (https://www.britannica.com/science/statistics/Residual-analysis#ref367525), were introduced into remote sensing literature by Ginevan (1979) and Aronoff (1982). The concepts were likely taken and re-phrased by some subsequent researchers in remote sensing studies as “producer's accuracy” and “user's accuracy”. Both Ginevan (1979) and Aronoff (1982) used the “producer's risk” to measure the probability of incorrectly rejecting an acceptable map, and the “consumer's risk” to measure the probability of accepting an inaccurate map. More specifically, Aronoff (1982) was addressing the classification accuracy of the binary (binomial) distribution in Alberta with two classes. However, he did not use nor recommend the terms “producer's accuracy” and “consumer's or user’s accuracy”. Aronoff (1982) actually used "proportion correct" and “% correct” to measure the accuracies in an error matrix. Accuracy and risk are two different concepts that should not be confused.

5. Often, when we use terms like “producer’s accuracy” and “user’s accuracy”, we have a tad of cynicism and sarcasm about the former. For instance, when we say:

   - Drug maker’s (producer’s) efficacy and patient’s (user’s) efficacy

   - Car maker’s (producer’s) fuel efficiency and car driver’s (user’s) fuel efficiency

   - Map maker’s (producer’s) accuracy and map user’s accuracy

   We often have a tad of cynicism and sarcasm about the producer’s (drug maker’s and car maker’s) claims. It is not difficult to infer that map maker’s (producer’s) accuracy and map user’s accuracy fall into the same “linguistic” category. Ideally we would prefer scientific terms with no extra linguistic connotations. We just want to be very clear on the correct proportions or percentages relative to the ground reference and the classification.

6. PR (correct Proportion or percentage relative to the Reference) and PC (correct Proportion or percentage relative to the Classification) in an error matrix are more intuitive and meaningful than the "producer's accuracy" and "user's accuracy". They are less prone to misunderstanding and more pertinent in representing the proportions (or percentages) of correct classifications in the error matrix. By calling the "producer's accuracy" and "user's accuracy" as PR and PC, respectively, we remove too many measures called "accuracies", which often have unique meaning in statistics and data science. We will not have a "100% user's accuracy" or a "100% perfect reliability" for a poor classification.

7. The notation "% Correct" was used by Aronoff (1982) to indicate the "correct percentage relative to the classification". He also used "% Commission" and "% Omission" to indicate commission and omission errors. However, Aronoff (1982) did not specifically use "% Correct" to indicate the "correct percentage relative to the ground reference", even though it is not difficult to infer such a term from his study in Alberta. We could have used "% Correct (reference)" or "% Correct_R" to indicate the correct percentage relative to the ground reference, and "% Correct (classification)" or "% Correct_C" to indicate the correct percentage relative to the classification (in fact, we used them in preliminary analyses). But they are considered not as concise as PR and PC.

8. The PR and PC are simply two correct proportions that correspond to the row and column for a category in an error matrix. They can be combined to provide a single measure (PAve) for each category (i.e., each species), by taking the pooled average of PR and PC. The PAve provides an overall classification performance measure for each category in an inventory. It is complementary to the "misclassification error" for each category (misclassification error for each category is the pooled average of omission error and commission error for each category). The PR, PC and PAve describe different aspects of a classification. They all have their roles to play in the classification. However, users may choose to focus on any one of the measures or balance all three measures for some intended purposes.

9. The PAve can be more meaningful in practice than either the PR or PC, if one is only interested in a singular overall accuracy for each individual species. It purports the reporting of a 100% accuracy or a 100% reliability only when it is a true 100% accuracy or reliability for a species in an inventory.

10. The PR and PC will always follow and be tied to the ground reference (in rows in our examples) and map classification (in columns in our examples). Switching the rows and columns in an error matrix will have no impact on the PR and PC values. We prefer to put "classification" from an aerial-based remote sensing technique on top (as it is typically obtained from above), and "reference" on level (as ground is usually used as the reference). This way the R in PR can also signify "row" and the C in PC can also signify column. But these preferences have no impact on the outcome.

While opinions and preferences may differ, and it could always be argued that there may not be anything fundamentally wrong with the conventional idioms (and that the new terms could be equally confusing if one is not careful), the vague, impertinent and easily misinterpretable nature of the conventional idioms made their practical use unwarranted, or at least undesirable. Most of us simply prefer the use of some terms that are more intuitive and less prone to misunderstanding and misinterpretation in practice. Some additional more technical reasons for our preference are imbedded in the discussion next.

5.2 Accuracy and agreement measures based on the error matrix

Many statistics can be calculated from an error matrix. To illustrate, an error matrix shown in Table 30 is used. It is a simplified version of the error matrix shown in Table 2. All variables in Table 30 are identical to those defined before for Table 2.

TABLE 30. An error matrix (in actual counts) with *k* possible categories.

		Classification				
	Category	1	2	⋯	*k*	**Total**
Reference	1	n_{11}	n_{12}	⋯	n_{1k}	nr_1
	2	n_{21}	n_{22}	⋯	n_{2k}	nr_2
	⋮	⋮	⋮	⋱	⋮	⋮
	k	n_{k1}	n_{k2}	⋯	n_{kk}	nr_k
	Total	nc_1	nc_2	⋯	nc_k	*N*

Note: nr_1, nr_2, ⋯, nr_k are row totals and nc_1, nc_2, ⋯, nc_k are column totals for categories 1, 2, ⋯, *k*, respectively, and *N* is the total number of counts.

To facilitate the discussion (later), Table 30 is also expressed in proportions by dividing the actual count in each cell by the total number of counts (*N*), producing the following corresponding table of proportions (Table 31).

TABLE 31. AN ERROR MATRIX (IN PROPORTIONS) WITH *K* POSSIBLE CATEGORIES.

		Classification				Total
	Category	1	2	⋯	*k*	
Reference	1	p_{11}	p_{12}	⋯	p_{1k}	pr_1
	2	p_{21}	p_{22}	⋯	p_{2k}	pr_2
	⋮	⋮	⋮	⋱	⋮	⋮
	k	p_{k1}	p_{k2}	⋯	p_{kk}	pr_k
	Total	pc_1	pc_2	⋯	pc_k	1

Note: this table corresponds to Table 30. It is obtained by dividing the count in each cell of Table 30 by the total number of counts (*N*).

Based on Table 30 or Table 31, a large number of accuracy and error measures can be calculated. Table 32 lists many of these measures, where i=1, 2, ⋯, k. While the efforts put into the development of these measures may be laudable, and there are numerous cases in which they were used in many remote sensing studies for different reasons and purposes, the practical utility and efficacy of most of these measures, however, are questionable and debatable to say the least.

TABLE 32 (PART 1 of 2). ACCURACY AND ERROR MEASURES DERIVED FROM AN ERROR MATRIX.

No.	Name	Formula or example reference
1	Overall accuracy (P_o)	$P_o = \sum_{i=1}^{k} n_{ii} / N = \sum_{i=1}^{k} p_{ii}$
2	Correct proportion re reference/row (PR)	$PR_i = n_{ii}/nr_i = p_{ii}/pr_i$
3	Correct proportion re classification/column (PC)	$PC_i = n_{ii}/nc_i = p_{ii}/pc_i$
4	Pooled average (PAve)	$PAve_i = 2n_{ii}/(nr_i+nc_i) = 2p_{ii}/(pr_i+pc_i)$
5	Overall error (E_o)	$E_o = 1 - P_o$
6	Omission error or error of omission (EO)	$EO_i = 1 - PR_i$
7	Commission error or error of commission (EC)	$EC_i = 1 - PC_i$
8	Pooled average error (PE)	$PE_i = 1 - PAve_i$
9	Kappa, kappa coefficient, or Cohen's kappa	Cohen 1960
10	Weighted kappa	Cohen 1968, Fleiss et al. 1969
11	Fleiss' kappa	Fleiss 1971
12	Kappa with random chance agreement	Foody 1992
13	Modified kappa	Liu et al. 2007
14	The KHAT statistic (an estimate of kappa)	Congalton et al. 1983
15	Scott's Pi	Scott 1955
16	Krippendorff's alpha coefficient	Krippendorff 1978, 2018
17	Producer's risk	Aronoff 1982
18	Aronoff's index of classification accuracy	Aronoff 1985
19	Combined accuracy (user's)	Nelson 1983
20	Combined accuracy (producer's)	Nelson 1983
21	Average accuracy (user's)	Fung and LeDrew 1988
22	Average accuracy (producer's)	Fung and LeDrew 1988
23	Average mutual information	Finn 1993
24	Map-level normalized accuracy	Congalton 1991, Zhuang et al. 1995
25	Normalized mutual information (NMI_{map})	Finn 1993
26	Normalized mutual information ($NMI_{reference}$)	Finn 1993
27	Normalized mutual information ($NMI_{map\ and\ reference}$)	Strehl and Ghosh 2002
28	The tau (τ) coefficient	Ma and Redmond 1995
29	Classification success index	Koukoulas and Blackburn 2001
30	Combined accuracy (user's and producer's)	Liu et al. 2007
31	Double average accuracy (user's and producer's)	Liu et al. 2007
32	Average of Short's mapping accuracy index	Liu et al. 2007
33	Average of Helldén's mean accuracy index	Liu et al. 2007
34	Overall agreement measure	Pontius and Santacruz 2014
35	Jaccard coefficient	Rosenfield and Fitzpatrick-Lins 1986
36	Ground truth index	Türk 1979
37	Short's mapping accuracy	Short 1982
38	Conditional kappa (map)	Rosenfield and Fitzpatrick-Lins 1986
39	Conditional kappa (reference)	Rosenfield and Fitzpatrick-Lins 1986
40	Category-level normalized accuracy	Congalton 1991
41	Relative change of entropy (map)	Finn 1993
42	Relative change of entropy (reference)	Finn 1993
43	Individual classification success index	Koukoulas and Blackburn 2001
44	Average of user's and producer's accuracy	Liu et al. 2007
45	Balanced accuracy	Chicco et al. 2021
46	F1 score	Chicco and Jurman 2020
47	Matthews correlation coefficient	Matthews 1975
48	Fowlkes–Mallows index	Fowlkes and Mallows 1983
49	Informedness or bookmaker informedness	Powers 2011

TABLE 32 (PART 2 of 2). ACCURACY AND ERROR MEASURES DERIVED FROM AN ERROR MATRIX.

No.	Name	Formula or example reference
50	Markedness	Powers 2011
51	Pearson's phi coefficient	Davenport and El-Sanhury 1991.
52	Diagnostic odds ratio	Tharwat 2020
53	Threat score or critical success index	Schaefer 1990

From a practical point of view, the first four measures in Table 32 are clear, succinct, useful and most relevant to the classification accuracy assessment (for categorical variables). Many other measures have limited or no value in the classification accuracy assessment. They mostly serve to obscure and muddy the information that is truly relevant and valuable to the question we want to answer. They are more distracting than enlightening to most practitioners. In fact, some of these measures are entirely wrong or misleading for accuracy assessment and comparison purposes. They actually create problems rather than solving them. Here, since a detailed comparison of these measures is not a main objective of this study, we only provide some brief explanations, notes, examples and summaries on most of these measures in Table 32. To facilitate the description and discussion, we again use the conventional terminologies for most of these measures as they appeared in the literature.

1. The Kappa Statistic

The kappa statistic has been widely used to evaluate the accuracy of classifications in hundreds or perhaps more of remote sensing studies. The controversies surrounding and the futilities of the kappa statistic are well-documented in remote sensing literature (Hudson and Ramm 1987, Stehman 1997, Olofsson et al. 2014, Stehman and Foody 2019, Foody 2020) and elsewhere (Feinstein and Cicchetti 1990, Powers 2012). It is known that the common practice of using the kappa statistic to indicate classification accuracy in remote sensing studies is flawed. We will only highlight a few key points here.

The kappa statistic can be calculated in two different ways, as in [5.1], where P_o is the overall accuracy and $P_e=\sum_{i=1}^{k} pr_i \cdot pc_i$ (notations for the proportions and actual numbers are defined in Tables 30-31):

[5.1] $$\kappa = \frac{P_o - P_e}{1 - P_e} = \frac{\sum_{i=1}^{k} p_{ii} - \sum_{i=1}^{k} pr_i \cdot pc_i}{1 - \sum_{i=1}^{k} pr_i \cdot pc_i} \quad \text{or} \quad \kappa = \frac{\sum_{i=1}^{k} n_{ii}/N - \sum_{i=1}^{k} \frac{nr_i}{N}\frac{nc_i}{N}}{1 - \sum_{i=1}^{k} \frac{nr_i}{N}\frac{nc_i}{N}} = \frac{N\sum_{i=1}^{k} n_{ii} - \sum_{i=1}^{k} nr_i \cdot nc_i}{N^2 - \sum_{i=1}^{k} nr_i \cdot nc_i}$$

To illustrate the calculation of the kappa statistic in [5.1], an error matrix shown in Table 33 is used. The kappa statistic calculated for the data in Table 33 is:

$$\kappa = \frac{\sum_{i=1}^{k} p_{ii} - \sum_{i=1}^{k} pr_i \cdot pc_i}{1 - \sum_{i=1}^{k} pr_i \cdot pc_i} = \frac{(37+11+16)/85 - (42\cdot52+18\cdot15+25\cdot18)/85^2}{1-(42\cdot52+18\cdot15+25\cdot18)/85^2} = 0.5869.$$

TABLE 33. AN ERROR MATRIX AND SOME RELATED CALCULATIONS.

		Classification			Total	PR
	Species	Sw	Sb	Fb	**(row)**	
	Sw	37	3	2	42	88% (37/42)
Ground (reference)	Sb	7	11	0	18	61% (11/18)
	Fb	8	1	16	25	64% (16/25)
Total (column)		52	15	18	85	
PC		71% (37/52)	73% (11/15)	89% (16/18)		P_o = 75%

Note: species are defined in Table 1. PR and PC are correct proportions relative to the reference (row) and classification (column), respectively.

The kappa statistic can range from -1 to 1. According to the conventional understanding, a value of κ=1 indicates that the classifications from two methods/raters/samples are in complete agreement. A value of κ=0 indicates that the classifications are no better than what would be expected by random chances. A value of κ>0 indicates that the classifications are better than what would be expected by random chances. A value of κ<0 indicates that the classifications are worse than random.

Evidently, part of the above conventional understanding is not correct. This is illustrated in Table 34 using five error matrices (cases), where for instance, the overall classification accuracy for two species is 99% (Case 5), yet the kappa statistic is negative (κ =-0.0068). It would be extremely untenable to argue that a 99% accuracy is "worse than random". In fact, for all overall accuracies in Table 34, no matter how good they are, the kappa statistics are all negative (κ<0), leading to "worse than random".

TABLE 34. EXAMPLE MATRICES AND CORRESPONDING OVERALL ACCURACY AND KAPPA VALUES.

		Classification									
		Case 1		Case 2		Case 3		Case 4		Case 5	
	Species	Sw	Sb	Sw	Sb	Sw	Sb	Sw	Sb	Sw	Sb
Reference	Sw	10	2	25	2	100	2	200	2	350	2
	Sb	3	0	3	0	3	0	3	0	3	0
Overall accuracy		67%		83%		95%		98%		99%	
Kappa statistic		-0.1905		-0.0870		-0.0234		-0.0118		-0.0068	
McNemar's test		0.2000		0.2000		0.2000		0.2000		0.2000	
(*p*-value)		(0.6547)		(0.6547)		(0.6547)		(0.6547)		(0.6547)	

Furthermore, the use of McNemar test as one of the potential alternatives in accuracy assessment is also of little or no value, as the McNemar test statistic (calculated as $\chi^2 = (b\text{-}c)^2/(b+c)$ for the 2×2 matrix $\begin{vmatrix} a & b \\ c & d \end{vmatrix}$), is a constant for all five cases in Table 34 (as it is independent of the *a* and *d* elements in the 2×2 matrix), regardless of how good or bad the overall classification accuracy is. McNemar test is a hypothesis test used on paired nominal data. It is very restrictive, and is limited to 2×2 matrices only. The suggestion to use it as a possible alternative for assessing classification accuracy (Foody 2020), is likely an unsuitable and unfitting option.

One of the particularly undesirable traits of the kappa statistic is that, in the conventional understanding, the kappa statistics were often used for comparing the quality of different classifications. Larger kappa values were thought to represent better classification accuracies. This conventional understanding is again not true. There are many cases in which a classification can be ranked as being more accurate than another in terms of the overall accuracy, but when it is ranked by the kappa statistic, a completely opposite order may occur. This is illustrated in Table 35 using two 2×2 matrices and three 3×3 matrices, where more accurate classifications ranked by the overall accuracy are completely reversed when ranked by the kappa statistic. This occurs for the 2×2 matrices as well as the 3×3 matrices, where the increasing overall accuracies correspond to the decreasing kappa statistics.

TABLE 35. ADDITIONAL EXAMPLE MATRICES AND CORRESPONDING OVERALL ACCURACY AND KAPPA VALUES.

	2×2 matrix				3×3 matrix								
Matrix	5	24	0	1	10	30	50	10	20	50	10	0	50
	0	3	2	27	0	0	10	0	0	10	0	0	10
					10	0	20	10	0	20	10	0	20
Overall accuracy	25%		90%		23%			25%			30%		
Kappa statistic	0.0376		-0.0465		-0.0484			-0.0588			-0.0938		

The trends in Table 34 and Table 35 illustrate the fact that larger kappa values could lead to opposite directions. They could indicate more accurate classifications, as in Table 34 where the overall accuracy increases (67% → 99%) as the kappa statistic becomes larger (-0.1905 → -0.0068). They could also indicate less accurate classifications, as in Table 35 where the increasing overall accuracies (e.g., 25% → 90%) are associated with the decreasing kappa values (e.g., 0.0376 → -0.0465). Clearly, this could be very dangerous and misleading in practical accuracy assessment if the kappa statistic is used.

Many other kappa-based statistics listed in Table 32 and frequently mentioned and used in remote sensing studies, such as weighted kappa, Fleiss' kappa, kappa with random chance agreement, modified kappa, conditional kappa and the KHAT statistic (an estimate of the kappa statistic), most likely suffer from similar problems illustrated above. In a study to evaluate the pixel-by-pixel agreement between two lodgepole pine site index maps derived from climate variables, we applied the weighted kappa statistic without enough diligence and forethought (Monserud et al. 2006). We were aware that there were hundreds of peer-reviewed research papers that had used the kappa or kappa-based statistics to judge the accuracy of classifications, so we thought that it must be correct to use them. Obviously we erred in our judgment and made a blunder in this regard. While the kappa and kappa-based statistics may still have some utilities in chance-based probability realms, or in certain specific areas related to the level of balance or symmetry of matrices, their continued use in practical accuracy assessment and in comparing different classifications from remote sensing studies, is incorrect and should be avoided.

2. Helldén's Mean Accuracy Index or Pooled Average

In its original description the mean accuracy index "denotes the probability that a randomly chosen point of a specific class on the map has a correspondence of the same class in the same position in the field and that a randomly chosen point in the field of the same class has a correspondence of the same class in the same position on the map" (Helldén 1980, p.18). This is a long definition that will likely cause some confusion to many practitioners. On the one hand, Helldén's mean accuracy index is regarded as "a logical (heuristic) development of Helldén and cannot be derived on either a probability basis or a mathematical

basis" (Rosenfield and Fitzpatrick-Lins 1986). On the other hand, it is just "the harmonic mean of producer's accuracy and user's accuracy" (Türk 2002, Liu et al. 2007).

The harmonic mean (HM) of a set of numbers x_1, x_2, ⋯, x_k is defined to be HM = $k/(1/x_1+1/x_2+\cdots+1/x_k)$. For instance, for the special case of three numbers, x_1, x_2 and x_3, the harmonic mean can be written as:

[5.2] $$\mathrm{HM} = \frac{3x_1x_2x_3}{x_1x_2+x_1x_3+x_2x_3}.$$

For two numbers, x_1 and x_2, the harmonic mean is:

[5.3] $$\mathrm{HM} = \frac{2x_1x_2}{x_1+x_2}.$$

The mean accuracy (MA) index used by Helldén (1980) for category i (i=1, 2, ⋯, k) is:

[5.4] $$\mathrm{MA}_i = \frac{2n_{ii}}{nr_i+nc_i} \quad \text{or} \quad \mathrm{MA}_i = \frac{2p_{ii}}{pr_i+pc_i}.$$

where n_{ii}, nr_i, nc_i, p_{ii}, pr_i and pc_i are defined in Tables 30 and 31. Since the "producer's accuracy" (PR_i) and the "user's accuracy" (PC_i) are defined by:

[5.5] $$\mathrm{PR}_i = n_{ii}/nr_i \qquad \mathrm{PC}_i = n_{ii}/nc_i$$

Their harmonic mean is (following [5.3]):

[5.6] $$\mathrm{HM} = \left(2 \times \frac{n_{ii}}{nr_i}\frac{n_{ii}}{nc_i}\right)\Big/\left(\frac{n_{ii}}{nr_i}+\frac{n_{ii}}{nc_i}\right) = \frac{2n_{ii}}{nr_i+nc_i} = \mathrm{MA}_i = \frac{2p_{ii}}{pr_i+pc_i}$$

So the mean accuracy index used by Helldén can in fact be readily derived on a mathematical basis. It is just "the harmonic mean of producer's accuracy and user's accuracy".

While the calculation of the harmonic mean is simple enough, the concept of "harmonic mean" is foreign to most practitioners in forestry. Indeed, the mean accuracy index can be more easily and simply understood mathematically or statistically as the pooled or weighted average of "producer's accuracy" and "user's accuracy", weighted by the sample sizes of nr_i and nc_i:

[5.7] $$\mathrm{PAve}_i = \frac{\mathrm{PR}_i\times nr_i+\mathrm{PC}_i\times nc_i}{nr_i+nc_i} = \frac{(n_{ii}/nr_i)\times nr_i+(n_{ii}/nc_i)\times nc_i}{nr_i+nc_i} = \frac{2n_{ii}}{nr_i+nc_i}.$$

As an example, for Sw in Table 33, the pooled average is PAve (Sw)=2×37/(42+52)=79%. For Sb and Fb, the pooled averages are PAve (Sb)=2×11/(18+15)=67% and PAve (Fb)=2×16/(25+18)=74%.

Most practitioners are very familiar with and have used for long the pooled or weighted average of two or more subsamples with different sample sizes. Therefore, it is intuitively more understandable and much simpler to use a term like "pooled average" or "weighted average", instead of the "harmonic mean" or "Helldén's mean accuracy index", even though in this study we still refer the term "Helldén's mean accuracy index" in the main text, to recognize the work of Helldén (1980), who used the pooled average. But there is really no need to evoke any probability theory in the definition and use of such a measure. It is simply a pooled average of the "producer's accuracy" (PR_i) and "user's accuracy" (PC_i).

3. Averaging the Averages

Many accuracy measures in conventional remote sensing studies were formulated by "averaging the averages". For example, given the overall accuracy ($P_o = \sum_{i=1}^{k} n_{ii}/N = \sum_{i=1}^{k} p_{ii}$), the "producer's accuracy" ($\mathrm{PR}_i = n_{ii}/nr_i = p_{ii}/pr_i$), the "user's accuracy" ($\mathrm{PC}_i = n_{ii}/nc_i = p_{ii}/pc_i$) and the pooled average ($\mathrm{PAve}_i = 2n_{ii}/(nr_i+nc_i) = 2p_{ii}/(pr_i+pc_i)$) defined earlier in Section 2.1 and Table 32, the following average accuracy measures (listed in Table 32) were formulated, used or compared in different remote sensing studies (Short 1982, Nelson 1983, Rosenfield and Fitzpatrick-Lins 1986, Fung and LeDrew 1988, Koukoulas and Blackburn 2001, Liu et al. 2007):

(1). **Average of user's and producer's accuracy** (AC_1)

[5.8] $\mathrm{AC}_1 = (\mathrm{PC}_i + \mathrm{PR}_i)/2$

(2). **Average accuracy - user's** (AC_2)

[5.9] $\mathrm{AC}_2 = \sum_{i=1}^{k} \mathrm{PC}_i/k$

(3). **Average accuracy - producer's** (AC_3)

[5.10] $AC_3 = \sum_{i=1}^{k} PR_i/k$

(4). **Double average accuracy - user's and producer's** (AC_4)

[5.11] $AC_4 = \frac{(\sum_{i=1}^{k} PC_i/k + \sum_{i=1}^{k} PR_i/k)}{2}$

(5). **Combined accuracy - user's** (AC_5)

[5.12] $AC_5 = (P_o + \sum_{i=1}^{k} PC_i/k)/2$

(6). **Combined accuracy - producer's** (AC_6)

[5.13] $AC_6 = (P_o + \sum_{i=1}^{k} PR_i/k)/2$

(7). **Average of Helldén's mean accuracy** (AC_7)

[5.14] $AC_7 = \sum_{i=1}^{k} PAve_i/k$

(8). **Combined accuracy - user's and producer's** (AC_8)

[5.15] $AC_8 = (P_o + \sum_{i=1}^{k} PAve_i/k)/2$

(9). **Individual classification success accuracy (index)** (AC_9)

[5.16] $AC_9 = PC_i + PR_i - 1$

(10). **Classification success accuracy (index)** (AC_{10})

[5.17] $AC_{10} = \sum_{i=1}^{k} PC_i/k + \sum_{i=1}^{k} PR_i/k - 1$

(11). **Short's mapping accuracy** (AC_{11})

[5.18] $AC_{11} = n_{ii}/(nr_i + nc_i - n_{ii})$ or $AC_{11} = p_{ii}/(pr_i + pc_i - p_{ii})$

The above average accuracy measures (AC_1-AC_{11}) are questionable to say the least. In fact, it was very surprising that they were included, used or compared as accuracy measures in remote sensing studies (Short 1982, Nelson 1983, Rosenfield and Fitzpatrick-Lins 1986, Fung and LeDrew 1988, Koukoulas and Blackburn 2001, Liu et al. 2007). They are not really meaningful in practice. To illustrate, assuming that an elementary school has two classes. If the average grade for class-1 with 30 students is 60, and the average grade for class-2 with 40 students is 90, the average grade for this school is (60×30+90×40) divided by the total number of students (30+40), or (60×30+90×40)/(30+40)=77.1429. If it was calculated as (60+90)/2 or (0.6+0.9-1), that would be completely wrong.

It is incorrect to simply average the averages to get a grand average when the samples sizes of different subsample groups or categories are different (this refers to measures (1)-(4) and (7)). In addition, we should not add a number (e.g., P_o, -1) to an average or averages, which can invalidate the results and interpretations (this refers to measures (5)-(6) and (8)-(10)). For instance, in calculating AC_5, one component, $\sum_{i=1}^{k} PC_i/k$, is already incorrect (as we should not do a simple average of the PC_i values when PC_1, PC_2, ⋯, PC_k for different categories are from different sample sizes). Adding another number (P_o) to an incorrect value would still produce an incorrect value. In calculating AC_{10}, two components, $\sum_{i=1}^{k} PC_i/k$ and $\sum_{i=1}^{k} PR_i/k$, are already incorrect as both PC_i and PR_i are from different sample sizes. Adding another number (which is "-1") to two incorrect values would still result in an incorrect value. Moreover, we should also avoid adding or subtracting a number (e.g., n_{ii} or p_{ii}) to the denominator of a fraction, which can twist the result and invalidate the interpretation (this refers to measure (11)). Since the denominator of a fraction typically represents the whole/total (a numerator of a fraction represents the number of parts out of the whole), artificially adding or subtracting a number to the denominator can be very misleading. It can distort the entire accuracy assessment and comparison outcome.

or instance, for the data in Table 33, following measure (11) expressed in [5.18], the AC_{11} values for Sw, Sb and Fb are calculated to be: AC_{11} (Sw) = 37/(42+52-37) = 65%, AC_{11} (Sb) = 11/(18+15-11) = 50%, and AC_{11} (Fb) = 16/(25+18-16)= 59%. These values, which are all smaller than the corresponding PR and PC values for the species, twist the accuracy assessment results and interpretations. What would be the meaning of a 50% for AC_{11} (Sb)? It would be a stretch to interpret it. On the other hand, regardless of one's preference in calling them, the PR, PC and PAve all have meaningful interpretations (e.g., the results in Table 33, P_o=(37+11+16)/85=75%, PR (Sb)= 61%, PC (Sb)=73% and PAve (Sb)=2×11/(18+15)=67%, can all be meaningfully interpreted).

While one could always "creatively define" a quantity to one's like and call it an accuracy measure, there are certain basic and fundamental mathematical and statistical rules that all researchers should all follow before conjuring up any accuracy measures. Otherwise, the so-called "accuracies" from these measures would, not only obscure and mask the real accuracy, but also create problems and lead to misleading inferences and erroneous conclusions.

As another example, for a classification (of a category) that is correct 10 out of 10 relative to the ground samples (PR_i=100%) and 10 out of 200 relative to the classifications (PC_i= 5%), the average for the category is (100%×10+5%×200)/(10+200)=9.5%. It would be wrong if it was calculated as (100%+5%)/2=52.5% following the "average of user's and producer's accuracy (AC_1)", or as (1+0.05-1)=5% following the "individual classification success accuracy (AC_9)", or as 10/(10+200-10)= 5% following the "Short's mapping accuracy (AC_{11})". The correct "average of averages" for subsamples having different sample sizes is the pooled average of subsample averages, defined in [5.19], where *k* is the number of subsample groups (e.g., categories):

[5.19] Pooled average = $\frac{Average_1 \times Samplesize_1 + Average_2 \times Samplesize_2 + \cdots + Average_k \times Samplesize_k}{Samplesize_1 + Samplesize_2 + \cdots + Samplesize_k}$

There is actually no need to use and evaluate any of the accuracies in [5.8]-[5.18]. It is just *plain wrong* to simply average the averages (many of which are proportions or ratios) without accounting for the numerators and denominators that the averages were derived from.

Furthermore, when so (too) many quantities are called "accuracy measures", there is really no accuracy measure any more (just like when we call too many things priorities, there is really no priority). The argument that these and many other accuracy measures listed in Table 32 might provide some unique information on certain unique aspects of a classification under some unique circumstances, was truly an unjustified stretch and a digression from the real accuracy assessment. Their uses and interpretations likely only serve to distract and confuse, rather than enlighten practitioners.

4. Normalized Accuracy and Error Measures

Table 32 lists several normalized accuracy and error measures (Congalton et al. 1983, Congalton 1991, Finn 1993, Zhuang et al. 1995, Strehl and Ghosh 2002). They have been referred or used in different applications (e.g., Ustin et al. 1996, Smits et al. 1999, Fashi et al. 2000, Liu et al. 2007). Typically, for instance, in the normalization to derive the normalized error matrix (Congalton et al. 1983), the original values in the error matrix must be "normalized" through "iterative proportional fitting", which forces each row and column in the matrix to sum to one. However, as pointed out by Stehman (1997), Stehman and Czaplewski (1998), Foody (2002), Stehman (2004) and Stehman and Foody (2019), normalization produces biased and imprecise accuracy estimates, with the bias most prominent for user's and producer's accuracies. It leads to estimates of parameters for a hypothetical population that has little relevance to the reality of the accuracy assessment.

The average mutual information (AMI) and its normalized forms have also been used as accuracy and error measures in the assessment of classification accuracy (Finn 1993, Strehl and Ghosh 2002, Liu et al. 2007). The AMI is calculated as:

[5.20] AMI = $\sum_{i,j=1}^{k} p_{ij} \ln\left(\frac{p_{ij}}{pr_i pc_j}\right)$ or AMI = $\sum_{i=1}^{k}\sum_{j=1}^{k} p_{ij} \ln\left(\frac{p_{ij}}{pr_i pc_j}\right)$

where p_{ij}s are the proportions defined in Table 31 (*i, j*=1, 2, ⋯, *k*), and "ln" denotes the *natural* (base $e \approx 2.71828$) *logarithm*. For instance, for the counts listed in Table 33 (*k*=3), the corresponding proportions (p_{ij}s) are listed in Table 36.

TABLE 36. AN ERROR MATRIX (IN PROPORTIONS) THAT CORRESPONDS TO TABLE 33 WITH THREE CATEGORIES.

		Classification			Total (row)
	Category/species	Sw	Sb	Fb	
Reference	Sw	p_{11}=37/85	p_{12}=3/85	p_{13}=2/85	pr_1=42/85
	Sb	p_{21}=7/85	p_{22}=11/85	p_{23}=0/85	pr_2=18/85
	Fb	p_{31}=8/85	p_{32}=1/85	p_{33}=16/85	pr_3=25/85
	Total (column)	pc_1=52/85	pc_2=15/85	pc_3=18/85	1 (85/85)

Note: species are defined in Table 1, pr_1, pr_2 and pr_3 are row total proportions and pc_1, pc_2 and pc_3 are column total proportions for categories 1, 2 and 3, respectively, and 85 is the total number of counts. More detailed definitions for the variables, numbers and proportions are provided in Tables 30, 31 and 33.

Based on the data in Table 36 and [5.20], the AMI is calculated to be:

[5.21] AMI = $p_{11}\ln(p_{11}/(pr_1 \times pc_1)) + p_{12}\ln(p_{12}/(pr_1 \times pc_2)) + p_{13}\ln(p_{13}/(pr_1 \times pc_3)) +$

$p_{21}\ln(p_{21}/(pr_2 \times pc_1)) + p_{22}\ln(p_{22}/(pr_2 \times pc_2)) + p_{23}\ln(p_{23}/(pr_2 \times pc_3)) +$

$$p_{31}\ln(p_{31}/(pr_3\times pc_1)) + p_{32}\ln(p_{32}/(pr_3\times pc_2)) + p_{33}\ln(p_{33}/(pr_3\times pc_3)) = 0.34487.$$

Therefore, following Finn (1993), Strehl and Ghosh (2002) and Liu et al. (2007), the normalized mutual information using the entropy on map (NMI_{map}), the normalized mutual information using the entropy on ground reference ($NMI_{reference}$) and the normalized mutual information using the (arithmetic) mean of the entropies on map and ground reference ($NMI_{map\ \&\ reference}$), can be computed:

[5.22] $$NMI_{map} = \frac{AMI}{-\sum_{i=1}^{k} pc_i \ln(pc_i)} = \frac{AMI}{-[pc_1\ln(pc_1)+pc_2\ln(pc_2)+pc_3\ln(pc_3)]} = \frac{0.34487}{0.93545} = 0.36867.$$

[5.23] $$NMI_{reference} = \frac{AMI}{-\sum_{i=1}^{k} pr_i \ln(pr_i)} = \frac{AMI}{-[pr_1\ln(pr_1)+pr_2\ln(pr_2)+pr_3\ln(pr_3)]} = \frac{0.34487}{1.03700} = 0.33256.$$

[5.24] $$NMI_{map\ \&\ reference} = \frac{2\times AMI}{[-\sum_{i=1}^{k} pc_i \ln(pc_i)]+[-\sum_{i=1}^{k} pr_i \ln(pr_i)]} = \frac{2\times 0.34487}{0.93545+1.03700} = 0.34969.$$

To illustrate the problems of AMI and its normalized forms without going into the details, following the above example, we calculated AMI, NMI_{map}, $NMI_{reference}$ and $NMI_{map\ \&\ reference}$ for five simpler error matrices listed in Table 37, each with only two categories or two species (Sw and Sb). Results of the calculations are listed in Table 37. For comparison, the overall accuracies and the kappa statistics associated with the five matrices are also calculated and listed in Table 37.

TABLE 37. EXAMPLE MATRICES AND CALCULATED AVERAGE MUTUAL INFORMATION AND ITS NORMALIZED FORMS.

		Classification									
		Case-I		Case-II		Case-III		Case-IV		Case-V	
	Species	Sw	Sb	Sw	Sb	Sw	Sb	Sw	Sb	Sw	Sb
Reference	Sw	10	2	200	2	5	24	0	1	25	0
	Sb	3	0	3	0	0	3	2	27	0	25
Overall accuracy (P_o)		67%		98%		25%		90%		100%	
Kappa statistic		-0.1905		-0.0118		0.0376		-0.0465		1.0	
AMI		0.03223		0.00014		0.01680		0.00234		0.69315	
NMI_{map}		0.08207		0.00263		0.03877		0.00956		1.0	
$NMI_{reference}$		0.06440		0.00189		0.05400		0.01601		1.0	
$NMI_{map\ \&\ reference}$		0.07217		0.00220		0.04514		0.01197		1.0	

Note: AMI=average mutual information, NMI_{map}=normalized mutual information (NMI) using the entropy on map, $NMI_{reference}$=NMI using the entropy on ground reference, and $NMI_{map\ \&\ reference}$=NMI using the (arithmetic) mean of the entropies on map and ground reference.

The futilities of the AMI, NMI_{map}, $NMI_{reference}$ and $NMI_{map\ \&\ reference}$ in indicating the classification accuracy should be clear from the results in Table 37. When the overall accuracy is 100% (case-V), the AMI= 0.69315. What does this AMI value mean? Does it mean a good or a poor classification accuracy? When the overall accuracy increases from 67% (case-I) to 98% (case-II) and 100% (case-V), the NMI_{map}, $NMI_{reference}$ and $NMI_{map\ \&\ reference}$ first decrease, then increase. For instance, the NMI_{map} first decreases from 0.08207 (case-I, P_o=67%) to 0.00263 (case-II, P_o=98%), then increases to 1.0 (case-V, P_o=100%). What do they mean for the classification accuracies in these cases? It would be highly untenable to use any of these and other "mutual information" related measures to determine the classification accuracy.

In probability theory and information theory, the mutual information or normalized mutual information of two random variables is a measure of the mutual dependence between the two variables (Kvålseth 2017). More specifically, it quantifies the "amount of information" between the two random variables in units foreign to most practitioners, such as "shannons (bits)", "nats" or "hartleys". It does not measure the correctness or accuracy (it measures "consistency" rather than "correctness" according to Finn (1993)). Using a combination of kappa and AMI to assess error matrices, as suggested by Finn (1993), is inapt and should be avoided in future studies as both kappa and AMI (and other AMI-induced measures such as relative change of entropy given a category on map and relative change of entropy given a category on ground reference) have little to do with the assessment of classification accuracy.

5. Other Accuracy and Error Measures

Throughout the main text (Sections 1-4), we purposely avoided mentioning "true positive", "true negative", "false positive", "false negative", "type I error" and "type II error" – terms and concepts that are near and dear to many statisticians' hearts but dreadful and confusing to many practitioners. Since these terms and concepts typically refer to the 2×2 (confusion) matrices in the context of accuracy assessment, and numerous accuracy and error measures are derived from the 2×2 matrices, we will just use the results in Table 4 (a 2×2 matrix), and put them into Table 38, to illustrate the concepts and calculations.

Table 38 lists the results of species classification from an inventory technique (lidar) for "two species", coniferous and non-coniferous (deciduous). Since there are only two species in Table 38, it belongs to the classic binary (two-class) classification,

where coniferous can be considered “positive” and non-coniferous can be considered “negative”. Hence, for other variables in Table 38, true positive (TP) refers to the correct classification for coniferous, true negative (TN) refers to the correct classification for non-coniferous, false positive (FP) or type-I error refers to the false alarm or overestimation of the positive, false negative (FN) or type-II error refers to the false alarm or overestimation of the negative, condition positive (P) is the number of actual positive counts in the data, condition negative (N) is the number of actual negative counts in the data, positive (PP) is the sum of true positive (TP) and false positive (FP) and negative (NN) is the sum of true negative (TN) and false negative (FN).

TABLE 38. A 2×2 CLASSIFICATION PERFORMANCE MATRIX FOR CONIFEROUS AND NON-CONIFEROUS (DECIDUOUS).

		Predicted (classification)		Total (row)
	Species	**Coniferous**	**Non-coniferous**	
Actual (ground)	Coniferous	True positive (TP) 119	False negative (FN) 7 (Type-II error)	Condition positive (P) 126
	Non-coniferous	False positive (FP) 2 (Type-I error)	True negative (TN) 81	Condition negative (N) 83
Total (column)		Positive (PP) 121	Negative (NN) 88	Grand total 209

Note: species are defined in Table 1. The data are from Table 4.

A large number of accuracy and error measures can be derived based on the variables defined in Table 38 (https://en.wikipedia.org/wiki/Confusion_matrix, Fawcett 2006, Powers 2011, Sammut and Webb 2011, Tharwat 2020, Chicco and Jurman 2020, Chicco et al. 2021). They include:

(1). **Balanced accuracy** (BA)

[5.25] $$BA = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right)$$

(2). **F1 score** (F1):

[5.26] $$F1 = \frac{2PPV \times TPR}{PPV+TPR} = \frac{2TP}{2TP+FP+FN}$$ ($TPR = \frac{TP}{P} = \frac{TP}{TP+FN}$; $PPV = \frac{TP}{TP+FP}$)

(3). **Matthews correlation coefficient** (MCC)

[5.27] $$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

(4). **Fowlkes–Mallows index** (FM)

[5.28] $$FM = \sqrt{\frac{TP}{TP+FP} \times \frac{TP}{TP+FN}}$$

(5). **Informedness or bookmaker informedness** (BM)

[5.29] $$BM = \frac{TP}{TP+FN} + \frac{TN}{TN+FP} - 1$$

(6). **Markedness** (MK)

[5.30] $$MK = \frac{TP}{TP+FP} + \frac{TN}{TN+FN} - 1$$

(7). **Threat score (TS) or critical success index**

[5.31] $$TS = \frac{TP}{TP+FN+FP}$$

These and other accuracy and error measures derived from the 2×2 confusion matrix have been used frequently in computer science and machine learning. Four of them, F1 score, bookmaker informedness (BM), markedness (MK) and Matthews correlation coefficient (MCC), have also appeared in remote sensing studies for accuracy assessment. Unfortunately, as we will see next, they should have no role to play in object-based accuracy assessment in remote sensing studies.

To explain the above statement, we again use the previous examples that showed the futility of the kappa statistic in remote sensing studies.

Table 39 lists five simple binary classification matrices, together with the calculated statistics defined in [5.25]-[5.31]. The overall accuracies and the kappa statistics for the matrices are also calculated and listed in Table 39.

TABLE 39. EXAMPLE MATRICES AND CORRESPONDING ACCURACY AND ERROR MEASURES.

		Classification									
		Case-I		Case-II		Case-III		Case-IV		Case-V	
	Species	Sw	Sb	Sw	Sb	Sw	Sb	Sw	Sb	Sw	Sb
Reference	Sw	10	2	200	2	5	24	0	1	24	1
	Sb	3	0	3	0	0	3	2	27	2	25
Overall accuracy (P_o)		67%		98%		25%		90%		94%	
Kappa statistic		-0.1905		-0.0118		0.0376		-0.0465		0.8846	
Balanced accuracy (BA)		0.4167		0.4951		0.5862		0.4655		0.9430	
F1 score (F1)		0.8000		0.9877		0.2941		0.0000		0.9412	
MCC		-0.1961		-0.0121		0.1384		-0.0496		0.8853	
Fowlkes–Mallows index (FM)		0.8006		0.9877		0.4152		0.0000		0.9414	
BM		-0.1667		-0.0099		0.1724		-0.0690		0.8859	
Markedness (MK)		-0.2308		-0.0148		0.1111		-0.0357		0.8846	
Threat score (TS)		0.6667		0.9756		0.1724		0.0000		0.8889	

Note: species are defined in Table 1, MCC=Matthews correlation coefficient, and BM=Bookmaker informedness.

It should be very clear from the results in Table 39 that the accuracy and error measures defined in [5.25]-[5.31] should have no role to play in judging the accuracy of any classifications from remote sensing studies. More specifically, for instance, when the overall accuracy increases from 25% (case-III) to 90% (case-IV) and 94% (case-V), the F1 score first decreases from 0.2941 to 0 as the P_o increases from 25% to 90%, then increases to 0.9412 as the P_o becomes 94%. The BA first decreases from 0.5862 (case-III, P_o=25%) to 0.4655 (case-IV, P_o=90%), then increases to 0.9430 (case-V, P_o=94%). The MCC first decreases from 0.1384 (P_o=25%) to -0.0496 (P_o=90%), then increases to 0.8853 (P_o=94%). The TS first decreases from 0.1724 (P_o=25%) to 0 (P_o=90%), then increases to 0.8889 (P_o=94%). Similar observations can also be made from other measures (FM, BM and MK). The trends displayed by these measures and the values themselves are as indefensible as the kappa statistic in judging the accuracy of classifications from remote sensing techniques. In fact, these measures may be useful in describing the balance or symmetry of the elements in a confusion matrix, but not in judging the accuracy of classifications unless certain theoretic restrictions and pre-conditions are met.

Many additional accuracy and error measures can also be formulated based on Table 38 (Fawcett 2006, Powers 2011, Tharwat 2020, Chicco et al. 2021). Some are listed in Table 40. Example calculations based on the TP, FN, FP and TN values in Table 38 are provided in Table 40 for interested readers.

TABLE 40. ADDITIONAL ACCURACY AND ERROR MEASURES DERIVED BASED ON TABLE 38.

Name	**Formula and example calculation (based on TP=119, FN=7, FP=2 and TN=81 from Table 38)**
True positive rate (TPR), hit rate or sensitivity	TPR = TP/P = TP/(TP+FN) = 0.9444
True negative rate (TNR), specificity or selectivity	TNR = TN/N = TN/(TN+FP) = 0.9759
Positive predictive value (PPV)	PPV = $\frac{\mathrm{TP}}{\mathrm{TP+FP}}$ = 0.9835
Negative predictive value (NPV)	NPV = $\frac{\mathrm{TN}}{\mathrm{TN+FN}}$ = 0.9205
False negative rate (FNR)	FNR = $\frac{\mathrm{FN}}{\mathrm{P}}$ = $\frac{\mathrm{FN}}{\mathrm{FN+TP}}$ = 0.05556
False positive rate (FPR) or fall-out	FPR = $\frac{\mathrm{FP}}{\mathrm{N}}$ = $\frac{\mathrm{FP}}{\mathrm{FP+TN}}$ = 0.02410
False discovery rate (FDR)	FDR = $\frac{\mathrm{FP}}{\mathrm{FP+TP}}$ = 0.01653
False omission rate (FOR)	FOR = $\frac{\mathrm{FN}}{\mathrm{FN+TN}}$ = 0.07955
Prevalence threshold (PT)	PT= $\frac{\sqrt{\mathrm{FPR}}}{\sqrt{\mathrm{TPR}}+\sqrt{\mathrm{FPR}}}$ = 0.1377
Positive likelihood ratio (PLR)	PLR = $\frac{\mathrm{TPR}}{\mathrm{FPR}}$ = 39.1944
Negative likelihood ratio (NLR)	NLR = $\frac{\mathrm{FNR}}{\mathrm{TNR}}$ = 0.05693
Diagnostic odds ratio (DOR)	DOR= PLR/NLR = 688.5

To facilitate the understanding and discussion next, the 2×2 confusion matrix shown in Table 38 is simplified to Table 41 below:

TABLE 41. AN EXAMPLE (BINARY) CLASSIFICATION MATRIX FOR TWO SPECIES (CONIFEROUS AND NON-CONIFEROUS).

		Predicted (classification)		Producer's accuracy
	Species	**Coniferous**	**Non-coniferous**	
Actual	Coniferous	True positive (TP)	False negative (FN)	TP/(TP+FN)
	Non-coniferous	False positive (FP)	True negative (TN)	TN/(FP+TN)
User's accuracy		TP/(TP+FP)	TN/(FN+TN)	

Based on Table 41 and the conventional terminologies associated with such a table, the "producer's accuracy" and "user's accuracy" can be calculated:

[5.32] Producer's accuracy = $\frac{\mathrm{TP}}{\mathrm{TP+FN}}$ for coniferous

[5.33] Producer's accuracy = $\frac{\mathrm{TN}}{\mathrm{FP+TN}}$ for non-coniferous

[5.34] User's accuracy = $\frac{\mathrm{TP}}{\mathrm{TP+FP}}$ for coniferous

[5.35] User's accuracy = $\frac{\mathrm{TN}}{\mathrm{FN+TN}}$ for non-coniferous.

Note that producer's accuracies for coniferous and non-coniferous are equivalent to TPR and TNR in Table 40, respectively; and user's accuracies for coniferous and non-coniferous are equivalent to PPV and NPV in Table 40, respectively.

However, in some remote sensing studies, the "producer's accuracy" and "user's accuracy" were completely mixed up (e.g., Scofield et al. 2015, Radoux and Bogaert 2017). "Producer's accuracy" was called "user's accuracy", and "user's accuracy" was called "producer's accuracy". For instance, Radoux and Bogaert (2017) expressed the "producer's accuracy" and "user's accuracy" as follows:

[5.36] Producer's accuracy = $\frac{\mathrm{TP}}{\mathrm{TP+FP}}$

[5.37] User's accuracy = $\frac{\mathrm{TP}}{\mathrm{TP+FN}}$

Radoux and Bogaert's (2017) "producer's accuracy" in [5.36] is actually the "user's accuracy" in [5.34], and their "user's accuracy" in [5.37] is actually the "producer's accuracy" in [5.32].

Researchers who used [5.36]-[5.37] or those that appeared in Scofield et al. (2015) likely misunderstood the original definitions of "producer's accuracy" and "user's accuracy", and flip-flopped the two terms. This is not surprising, for the concepts of "confusion matrix", "producer's accuracy", and "user's accuracy" or "reliability" can be very confusing and easily mixed up. This is also a reason that why practitioners may need to avoid or abandon these concepts and the accompanying measures in assessing classification accuracy. Radoux and Bogaert (2017) also proposed several other related measures, which we will not comment here. Other researchers also attempted to expand some of the measures in [5.25]-[5.31], which apply to binary classifications only, to multi-class classifications (e.g., Matthews correlation coefficient in [5.27] has been generalized to multi-class classifications, https://en.wikipedia.org/wiki/Matthews_ correlation_coefficient). We will not comment on them either, but caution that, they are kappa-like statistics and for assessing classification accuracy in remote sensing studies, they do not even apply to the simplest two-class classifications (see Table 39).

While the efforts to create new or introduce existing concepts and ideas from other scientific disciplines into remote sensing studies are commendable, past experience and literature suggested that, in accuracy assessment, it could be very challenging. Misunderstanding and misuses occurred frequently in published literature. This was most clearly evident from the widely and inappropriately adopted kappa type of statistics in remote sensing studies. The introduction of many accuracy and error measures derived from a confusion matrix, such as those defined in [5.25]-[5.31] and Table 40 and frequently seen in computer science and machine learning, provides no further improvement over the kappa statistics when used in judging the accuracy of classifications in remote sensing studies (e.g., see Table 39). Like the kappa statistics, most of these measures derived from a confusion matrix may have some utilities in very theoretical chance-based probability realms in machine learning and computer science, or in certain specific areas related to the level of balance or symmetry/asymmetry of the confusion matrix, their use in object-based accuracy assessment and in comparing different classifications from remote sensing techniques, should be avoided. Otherwise, we could make other kappa-sized debacles.

Practitioners should be particularly aware that there are numerous idiosyncratic accuracy and error measures that can be derived from a "confusion matrix" as it is termed in machine learning and computer science. Many of these measures bear a nebulously labelled moniker like "accuracy index", "truth index", "success index", "average mutual information", "normalized

mutual information", "entropy" or "relative change of entropy" (e.g., see Table 32). Moreover, there seems to be no agreement in the research community on which measures are better. It could always be argued from a theoretical point of view that, each of these measures might provide a different piece of information contained in the confusion matrix, and each might be more relevant or unique than others in a particular situation, under a particular circumstance, for a particular study with a particular goal and objective, or for a different theme, etc.

However, the use of these different accuracy measures frequently results in different, inconsistent and conflicting interpretations and conclusions. To most practitioners, they are more distracting than enlightening. More importantly, most of these measures and their interpretations are of no real practical value in indicating classification accuracy and comparing different classifications in remote sensing studies. They only serve to obscure, muddy, distract and unnecessarily complicate the accuracy information that is truly relevant and valuable to the question we want to answer. Many of these measures are entirely irrelevant, misleading or completely wrong for assessing and comparing classification accuracies when applied in remote sensing studies.

A Note on the Definition of Remote Sensing

In this study, we used several terms to describe an inventory or an inventory technique, such as "aerial-based inventory", "remote sensing-based inventory", "aerial-based remote sensing technique", etc. We are still unsure about the proper definition of "remote sensing". Traditionally, for instance, these definitions of remote sensing are commonly used:

1. From Lillesand et al. (2015):

   > "Remote sensing is the science and art of obtaining information about an object, area, or phenomenon through the analysis of data acquired by a device that is not in contact with the object, area or phenomena under investigation."

2. From U.S. Geological Survey (https://www.usgs.gov/faqs/what-remote-sensing-and-what-it-used#):

   > "Remote sensing is the process of detecting and monitoring the physical characteristics of an area by measuring its reflected and emitted radiation at a distance (typically from satellite or aircraft)."

3. From U.S. National Oceanic and Atmospheric Administration (NOAA) (https://oceanservice.noaa.gov /facts/remotesensing.html):

   > "Remote sensing is the science of obtaining information about objects or areas from a distance, typically from aircraft or satellites."

4. From Wikipedia (https://en.wikipedia.org/wiki/Remote_sensing):

   > "Remote sensing is the acquisition of information about an object or phenomenon without making physical contact with the object, in contrast to in situ or on-site observation."

Although "remote sensing" is defined in different ways in the above examples, the essence embedded in these definitions remains the same: it is the acquisition of information about an object from a distance, without making physical contact with the object.

In forest mensuration and forest inventory, information about tree and stand attributes is frequently obtained from a distance without making physical contact with the trees in a stand. This is regularly done through "point sampling", "point plots", "prism sweeps", "prism plots" or "variable-radius plots", based on, for example, different "basal area factors" (BAFs). If the above remote sensing definitions were proper, every time when we measure a plot using prism, it would be considered applying remote sensing.

Obviously, the vast majority of practitioners would not consider measuring a plot using prism to be "remote sensing", even through some academics argued in the research realm that as soon as one opens the eyes and starts to see or read something, one is employing "remote sensing" (Lillesand et al. 2015). We will not contest this here but note that many of the "remote sensing" definitions were proposed by engineers, who may not be aware of the forest inventory techniques we use regularly (like prism sweeps and BAFs). In fact, remote sensing is likely better to be defined as follows, at least in forestry:

- Remote sensing is the science of obtaining information about objects or areas on the ground from the air or space, via aircrafts, satellites, or other aerial-based vehicles, devices or sensors on different platforms.

Although for some other reasons, we refrained from calling it "aerial sensing" in this study, the above definition may be more fitting as it connotes the use of air- or space-based sensors and technologies to acquire information on the ground.

Practitioners typically do not call ground-based prism sweeps or tri-pod mounted devices or sensors remote sensing, although this might still be debatable among some researchers and academics. We will leave this up to the readers to ponder and explore further.

5.3 Agreement measures for continuous variables

Some of the accuracy measures listed in Table 32 for categorical variables were frequently referred to as agreement measures. In fact, the original kappa statistic put forward by Cohen (1960) was termed "a coefficient of agreement". Subsequent modifications to the kappa statistic were also frequently referred to as agreement measures. It appears that the terms "accuracy measures" and "agreement measures" have been used interchangeably for categorical variables in many remote sensing studies.

For continuous variables, the traditional goodness-of-fit statistics such as those presented in Section 4.1 and the coefficient of determination (R^2) and Pearson's correlation coefficient (r), are typically related to the correlation (association) between two variables, observed and predicted, where "observed" can mean observed on the ground or from an accepted reference standard, and "predicted" can mean classified, extracted, interpreted, estimated or mapped, etc. In day-to-day usages, the correlation between two variables is often mixed with the agreement between two variables. Fundamentally, though, correlation and agreement are two entirely different concepts (Robinson 1957; Altman and Bland 1983, 1987; Bland and Altman 1986, Hollis 1996; Stehman 1997; Liao and Lewis 2000, Ludbrook 2002, Bunce 2009, Agresti 2013; Choudhary and Nagaraja 2017). This is explained in detail in Huang et al. (2019) using forestry examples. For accuracy assessment and comparison, we could look at both correlation and agreement, but the focus must be on accuracy and agreement.

A large number of agreement measures under various names, such as "agreement index", "intra- and inter-class correlation coefficients", "agreement coefficient", "measure of agreement" and "concordance correlation coefficient" have been proposed by different researchers across different scientific disciplines. Comprehensive reviews of these measures focused primarily on their uses in clinical trials, pharmacology and other medical fields are provided by many researchers (Müller and Büttner 1994, Atkinson and Nevill 1998, Krummenauer and Doll 2000, Choudhary and Nagaraja 2005, Liao and Capen 2009, Ungerer and Pretorius 2017, Barnhart 2018). At least eight books, to varying degrees of scope and depth, focusing almost entirely on various aspects of agreement measures and agreement evaluations have been published within the last 18 years (Dunn 2004, von Eye and Mun 2004, Broemeling 2009, Carstensen 2010, Shoukri 2010, Lin et al. 2012, Gwet 2012, Choudhary and Nagaraja 2017).

Table 42 lists nine of the more commonly seen agreement measures. The index of agreement (d) and agreement coefficient (AC) appear to have been used relatively more frequently in remote sensing and some other scientific disciplines (e.g., Willmott et al. 1985, Legates and McCabe 1999, Ji et al. 2008, Riemann et al. 2010, Wilson et al. 2012, Zald et al. 2016, Duveiller et al. 2016, Neeti and Kennedy 2016, Matasci et al. 2018).

TABLE 42. AGREEMENT MEASURES FOR CONTINUOUS VARIABLES.

No.	Name	Reference
1	Coefficient of agreement (COA)	Robinson 1957
2	Measure of agreement (MOA)	Mielke 1984, 1991
3	Modified measure of agreement (M)	Watterson 1996
4	Concordance correlation coefficient (CCC_1)	Lin 1989, Lin et al. 2002
5	Improved CCC (CCC_2)	Liao 2003
6	Agreement parameter (λ)	Duveiller et al. 2016
7	Index of agreement (d)	Willmott 1981, 1982
8	Agreement coefficient (AC)	Ji and Gallo 2006
9	Limits of agreement (LoA)	Altman and Bland 1983, Bland and Altman 1986

Comparison of the agreement measures listed in Table 42 is not a trivial task. Interested readers can read Huang et al. (2019) for some technical details and lengthy analyses and discussions (including those on why the R^2 and r are not relevant in agreement analysis). Only some of the relevant highlights are listed here.

1. Many agreement measures are similar. Several are equivalent mathematically, or are simple re-scaling of or modification to each other. They appear logically similar in interpretations to the traditional R^2 and r statistics, but they were designed to be better and more meaningful than the R^2 and r in agreement studies. It was shown that the measure of agreement (MOA), modified measure of agreement, concordance correlation coefficient (CCC_1), improved concordance correlation coefficient and agreement parameter (λ) all functioned similarly. In fact, CCC_1 is identical to MOA, and λ is identical to CCC_1 and thus to MOA in all cases in agreement studies where $r \geq 0$. Thus, for practical uses, Mielke's MOA is recommended for its originality and theoretical soundness.

2. The index of agreement (*d*) proposed by Willmott (1981) has been used widely across different disciplines (e.g., Duveiller et al. 2016, Neeti and Kennedy 2016 (with a typo in the *d* formula)). It was originally designed for model validation, where one set of measurements (y_i) was "fixed" and considered to be the truth or "gold standard", and the other set of measurements (x_i) was model predictions to be compared to the truth. For agreement studies, whose main purpose is to assess if the values from two measurements are comparable and if one measurement can be replaced by the other more efficient measurement or model prediction, the *d* possesses a critical drawback. It is not invariant to the switching of the y_i and x_i values, which means that the *d* values will be different or inconsistent if the positions of *y* and *x* are switched in the calculation of *d*.

   For instance, for a simple data set of three *y*-*x* pairs used in the calculation of MOA (see Section 4.1), y_i = 4, 12, 18 and x_i = 3, 20, 21, the *d* is 0.8765 (from $d = 1 - \sum(y_i - x_i)^2/(\sum(|y_i - \bar{y}| + |x_i - \bar{y}|)^2)$, see Willmott 1981, 1982).

   However, if the positions of y_i and x_i are switched, i.e., if x_i = 4, 12, 18 and y_i = 3, 20, 21, the *d* becomes 0.8872 (as opposed to 0.8765). This essentially implies that if *d* is used as an agreement measure, the agreement can be different between the same sets of data. This would be analogous to argue, for instance, that two people A and B agree with each other, and from A's point of view this is true. But from B's point of view this may not be true. Obviously as an agreement measure this critical drawback (inconsistency) would be considered undesirable.

3. The agreement coefficient (AC) proposed by Ji and Gallo (2006) has some distinct shortcomings that make its utility inadvisable and unwarranted. For instance, Ji and Gallo (2006) claimed that "AC is bounded below by 0 and above by 1" (i.e., $0 \leq AC \leq 1$), which is not true. The AC is not only unbounded by zero, but also not bounded by -1, i.e., it is out of bounds on the negative side. More strikingly, the AC could produce completely opposite values in many cases to what is supposed to be and what the real agreement is. Positive agreement could become negative agreement and vice versa.

   To illustrate the seriousness of the problems associated with the AC, a simple data set of three positively related *y*-*x* pairs is used, y_i = [18.0, 23.0, 30.0] and x_i = [22.0, 24.0, 24.5]. The calculated AC for this data set is AC = -1.645, from $AC = 1 - \sum(y_i - x_i)^2/(\sum(|\bar{y} - \bar{x}| + |y_i - \bar{y}|)(|\bar{y} - \bar{x}| + |x_i - \bar{x}|))$. It is obvious that this AC = -1.645 is not bounded by 0 (not even by -1). Moreover, it is negative for an obvious positive relationship between *y* and *x* (i.e., *y* increases as *x* increases, with *r* = 0.909 between y_i and x_i). A completely wrong conclusion would have been reached if AC was used and interpreted as the agreement measure.

   To make matters worse, the errors associated with the calculation of the AC also propagate into the calculation of the so-called unsystematic (AC_u) and systematic (AC_s) agreement coefficients. This makes any inferences based on the AC_u and AC_s (and AC) inappropriate. For the three positively related *y*-*x* pairs (y_i = [18.0, 23.0, 30.0] and x_i = [22.0, 24.0, 24.5]), the AC_u and AC_s were calculated to be (see Huang et al. 2019, p.54), AC_u = 0.838 and AC_s = -1.483. The negative AC_s and subsequently the AC_u are totally meaningless.

   Fundamentally, the problems associated with the AC, AC_u and AC_s stem from the questionable definitions of AC, AC_u and AC_s, rendering any inferences based on them irrelevant and misleading. Note that this occurred when the positive trend in the data set is quite obvious and strong (*r* = 0.909).

   Irrational "flip-flop" AC, AC_u and AC_s values were also observed on many other data sets (examples of such data sets are available to interested readers). Similar observations were also reported by Duveiller et al. (2016), who demonstrated the AC problems through theoretical considerations, simulations and actual data. Indeed, it is very dangerous to use the AC and AC-induced measures to make any inferences when measuring and comparing the agreement between any two sets of measurements from different methods, devices or techniques, as they can lead to distorted, irrelevant or completely opposite results to what the correct results should be. It is really indefensible and injudicious that some remote sensing studies used the AC and AC-induced measures as their ultimate accuracy and agreement measures (see examples in Duveiller et al. 2016 and Huang et al. 2019).

4. The "limit of agreement" (LoA) concept introduced by Bland and Altman (1986) represents a real breakthrough in agreement analysis. It has shifted the paradigm and transformed how agreement analysis (and method comparison studies in general) should be conducted. The LoA-based Bland-Altman analysis is the most common and standard analytic technique used for agreement studies across a wide range of scientific disciplines. Bland and Altman's (1986) paper, based on a well-established and widely accepted "very simple and obvious" concept, and by virtue of its simplicity and readability, was listed by the leading scientific journal *Nature* as one of the top 100 most-referenced research papers of all time (Van Noorden et al. 2014), with 48,778 citations to date (May 19, 2021).

The LoA (or Bland-Altman analysis, Bland-Altman plot) has a number of distinct advantages, including: 1). It possesses the good characteristics of simplicity and validity. Only some well-established, widely used very simple statistics familiar to most practitioners need to be computed; 2). It avoids the problematic (and often irrelevant and misleading) and complex agreement measures; 3). It can be clearly decomposed into separate bias and precision components to provide a clear and complete depiction about bias, precision and accuracy; 4). It emphasizes the use of simply and neatly presented, easy-to-understand graphics, which are more informative, illuminating and powerful than many complex statistics; 5). It is directly linked to the same scale as the original data, allowing for easy interpretation on the same scale as the original data.

It is for these advantages of LoA, plus the above highlights and discussions on other agreement measures especially the index of agreement and agreement coefficient, that we recommended and used the LoA and MOA in our analysis, together with other statistics and statistical methods described in Sections 2, 3 and 4, as the methods for determining the accuracy of classifications and comparing the agreement between ground measures and forest inventory techniques.

5.4 Caveats on the chi-square test and Kolmogorov-Smirnov test

Earlier in Section 3.4, it was stated that the chi-square test and the KS test only enable us to determine whether there is a statistically significant difference between two sets of frequencies in proportions, but not two sets of frequencies in actual numbers, nor two sets of vaguely defined distributions. In short, the chi-square test and the KS test only test the frequency proportions, but not frequency numbers.

However, a casual search of the literature would lead to many articles that claim that such a test could be used to test, e.g., the null hypothesis (H_0) that the distributions from two samples are the same, against the alternative hypothesis (H_a) that the distributions from two samples are not the same. Clearly there are some confusion and probable misunderstanding in this area either way, especially about the KS test.

To avoid dwelling on some potentially futile argument on too many technical details, and to facilitate the understanding by most practitioners in our subject areas, we will just use two examples to illustrate the point. One for the chi-square test. The other for the KS test. Since Fisher's exact test functions the same as the chi-square test (but is usually considered better for small sample sizes), the logic imbedded in the example for the chi-square test also applies to Fisher's exact test for our purpose.

The Chi-Square Test

The chi-square test applies to categorical variables. Table 43 lists the species frequency counts for the same area from the ground and two inventory techniques, inventory-1 and inventory-2. The frequency counts from the ground and inventory-1 are actual observations taken directly from Table 3. Inventory-2 is assumed to be a completely wrong inventory (inventory-2 counts = 19 times inventory-1 counts). The frequency proportions (in %) correspond to the frequency counts are listed in Table 43. The means and the standard deviations (SD) of the frequency counts are also listed in Table 43. Note that the frequency proportions for inventory-1 and inventory-2 are identical for each species, but the frequency counts for inventory-2 are 19 times of those for inventory-1.

TABLE 43. SPECIES FREQUENCY COUNTS FROM THE GROUND AND TWO INVENTORY TECHNIQUES.

Type	Aw	Bw	Dp	Fb	Lt	Pb	Pl	Sb	Sg	Sw	Total	Mean	SD
Ground	55	18	0	8	14	10	9	28	4	63	209	20.9	21.6
Proportion (in %)	26.3	8.6	0.0	3.8	6.7	4.8	4.3	13.4	1.9	30.1	100		
Inventory-1	48	15	0	6	9	7	6	18	2	45	156	15.6	17.2
Proportion (in %)	30.8	9.6	0.0	3.8	5.8	4.5	3.8	11.5	1.3	28.8	100		
Inventory-2	912	285	0	114	171	133	114	342	38	855	2964	296.4	326.1
Proportion (in %)	30.8	9.6	0.0	3.8	5.8	4.5	3.8	11.5	1.3	28.8	100		

Note: SD is the standard deviation. Inventory 1 and Inventory 2 represent two inventory techniques. Species are defined in Table 1.

The frequency counts in Table 43 are displayed in Figure 15, where Figure 15(a) shows the frequency counts from the ground and inventory-1 and Figure 15(b) shows the frequency counts from the ground and inventory-2. It is strikingly clear from Figure 15(b) that the frequency counts from the ground and inventory-2 are very different, because inventory-2 is from a completely wrong inventory. The frequency distribution of inventory-2 (in terms of the mean and SD of the actual counts) is entirely different from those of the ground and inventory-1, as shown in Table 43.

Table 44 shows the computations for the chi-square test between the ground and inventory-2 (similar computations between the ground and inventory-1 were shown in Table 11). The calculated chi-square statistic is $\chi^2 = 3.1913$, which is the sum of the

last two columns of Table 44. This χ^2 = 3.1913 statistic corresponds to a *p*-value of 0.9218, which is greater than α = 0.05. This means that there is no significant difference between the two sets of frequency proportions in Table 43 for the species from the ground and inventory-2.

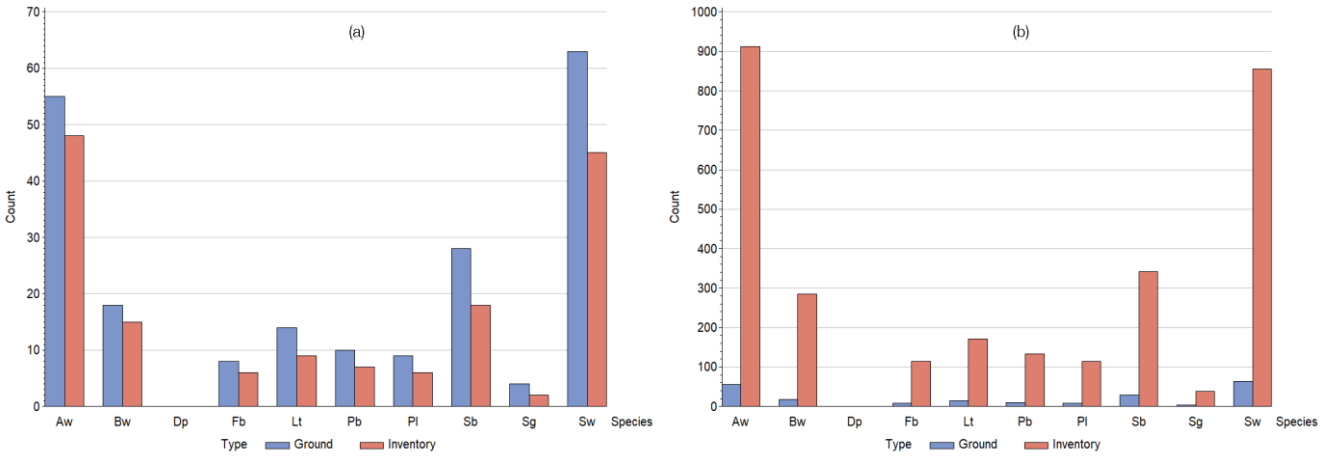

Figure 15. Ground frequency counts compared to the frequency counts from inventory-1 (a) and inventory-2 (b). The data are listed in Table 43.

It would be completely wrong if the chi-square statistic was interpreted to mean that there was no significant difference between the two sets of frequency numbers in Table 43 for the species from the ground and inventory-2. The frequency numbers for the species from the ground and inventory-2 are entirely different (inventory-2 is from a completely wrong inventory). They have drastically different means and SDs (mean=20.9 and SD=21.6 for the ground data; and mean=296.4 and SD=236.1 for inventory-2 classifications). It would be totally incorrect to mix frequency proportions with frequency numbers (counts) when using the chi-square test for inferencing.

Table 44. The chi-square test for the data from the ground and inventory-2.

Species	Ground (O_1)	Inventory 2 (O_2)	Row total	E_1	E_2	$(O_1-E_1)^2/E_1$	$(O_2-E_2)^2/E_2$
Aw	55	912	967	63.69	903.31	1.1869	0.0837
Bw	18	285	303	19.96	283.04	0.1921	0.0135
Fb	8	114	122	8.04	113.96	0.0002	0.0000
Lt	14	171	185	12.19	172.81	0.2702	0.0190
Pb	10	133	143	9.42	133.58	0.0358	0.0025
Pl	9	114	123	8.10	114.90	0.0996	0.0070
Sb	28	342	370	24.37	345.63	0.5403	0.0381
Sg	4	38	42	2.77	39.23	0.5500	0.0388
Sw	63	855	918	60.47	857.53	0.1061	0.0075
Total	209	2964	3173	209.00	2964.00	2.9811	0.2102

Note: the calculations follow [3.1] and [3.2], where the number of species (rows) is *i*=1, 2, …, 9 and the number of columns is *j*=1, 2. The ground and inventory-2 counts are listed in Table 43, O_1 and O_2 are observed values and E_1 and E_2 are corresponding expected values.

Hence, it was emphasized earlier in Section 3.4 that it was critically important to recognize that the chi-square test only enables us to determine if there is a significant difference between two frequency proportions, but not two frequency numbers, nor two distributions unless the "distributions" are explicitly defined as the frequency proportions. The chi-square test only allows us to test whether two sets of proportions from two samples for a categorical variable differ from each other. It does not allow us to test whether two sets of actual numbers from two samples for a categorical variable differ from each other.

The Kolmogorov-Smirnov Test

The KS test is primarily used for continuous variables. Table 45 lists the tree height frequency counts for a species from the ground and two inventory techniques, inventory-1 and inventory-2. The counts from the ground and inventory-1 are identical to those listed in Table 17 and shown in Figure 5. Inventory-2 is assumed to be a completely wrong inventory (inventory-2 counts = 21 times inventory-1 counts). The cumulative counts and the cumulative proportions that correspond to the frequency counts are also listed in Table 45. Note that the cumulative proportions for inventory-1 and inventory-2 are identical (but the cumulative counts for inventory-2 are 21 times of those for inventory-1).

TABLE 45. TREE HEIGHT FREQUENCY COUNTS FROM THE GROUND AND TWO INVENTORY TECHNIQUES.

Type	Tree height (m)								Total
	12	13	14	15	16	17	18	19	
Ground	1	2	0	1	0	4	3	1	12
Cumulative count	1	3	3	4	4	8	11	12	
Cumulative proportion	0.0833	0.25	0.25	0.3333	0.3333	0.6667	0.9167	1.0000	
Inventory-1	2	1	2	0	1	3	1	0	10
Cumulative count	2	3	5	5	6	9	10	10	
Cumulative proportion	0.2	0.3	0.5	0.5	0.6	0.9	1.0	1.0	
Inventory-2	42	21	42	0	21	63	21	0	210
Cumulative count	42	63	105	105	126	189	210	210	
Cumulative proportion	0.2	0.3	0.5	0.5	0.6	0.9	1.0	1.0	

The data in Table 45 are displayed in Figure 16, where the top two graphs show the ground frequency counts compared to the frequency counts from inventory-1 (a) and inventory-2 (b), and the bottom two graphs show the cumulative proportions from the ground compared to inventory-1 (c) and inventory-2 (d). Notice that the bottom two graphs are identical (even though the cumulative counts for inventory-1 and inventory-2 are drastically different).

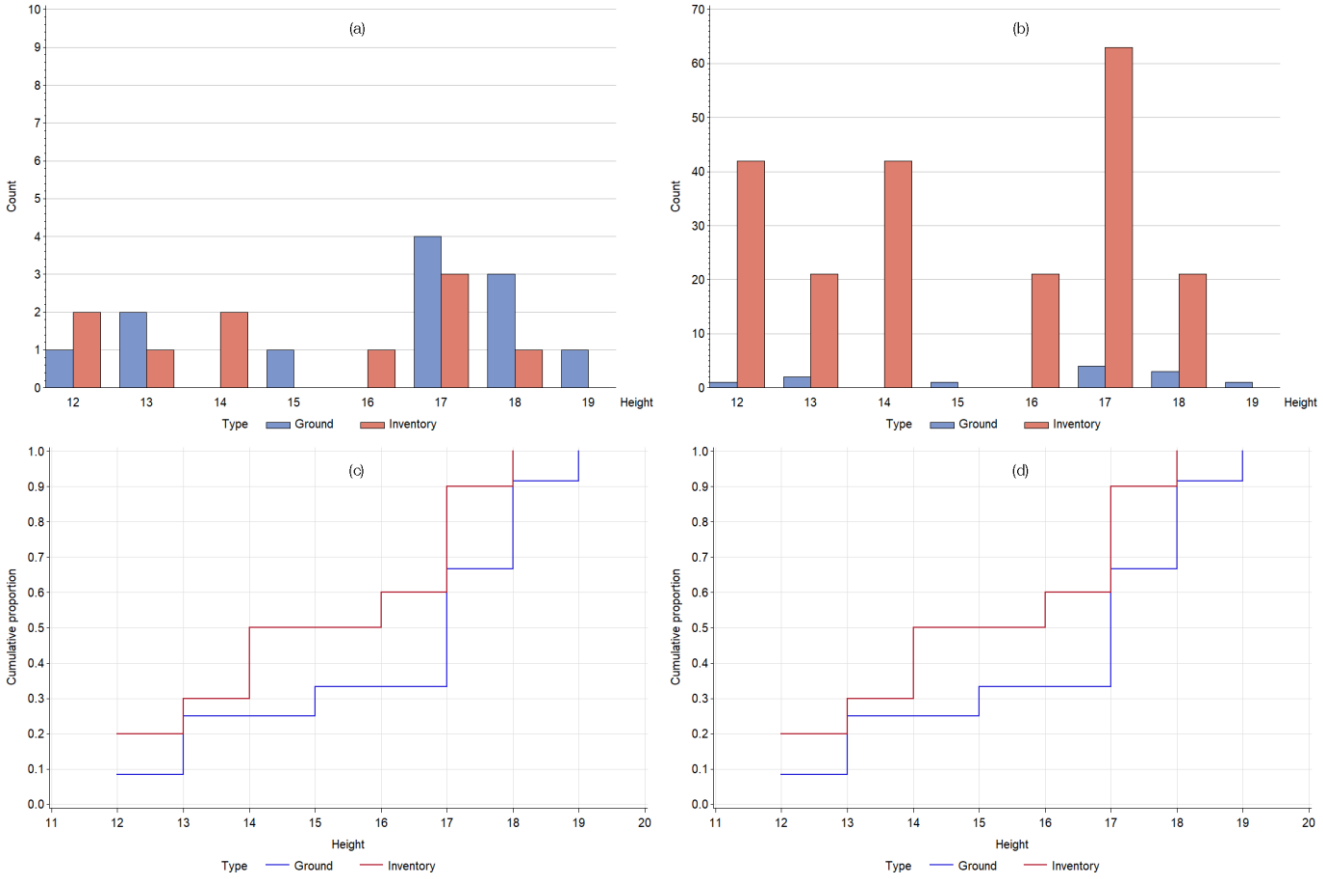

Figure 16. Top graphs show ground frequency counts compared to the frequency counts from inventory-1 (a) and inventory-2 (b). Bottom graphs show cumulative proportions from the ground compared to those from inventory-1 (c) and inventory-2 (d). Actual data are listed in Table 45.

It is rather clear from Figure 16(b) that the frequency counts from the ground and inventory-2 are very different (on average the frequency counts of inventory-2 are 210/12=17.5 times of those from the ground). This is not surprising because inventory-2 is from a completely wrong inventory.

Calculations of the KS test statistic (the *D* statistic) for the frequency counts from the ground and inventory 2 follow those demonstrated in Table 18 from the ground and inventory-1. The calculated *D* (*D*=0.26667) is identical to that shown in Table 18 (readers can actually infer this from the identical graphs in Figure 16(c) and Figure 16(d)). Therefore, the *z* calculated according to [4.12] is:

$$z = D\sqrt{\frac{n_1 n_2}{n}} = 0.26667\sqrt{\frac{12\times 210}{222}} = 0.89846.$$

Hence, the *p*-value computed by [4.11] or [4.13] is:

$$p\text{-value} = 2[(-1)^{(1-1)}e^{(-2z^2)} + (-1)^{(2-1)}e^{(-2\times 2^2 z^2)} + (-1)^{(3-1)}e^{(-2\times 3^2 z^2)}] + \cdots$$

$$= 2[0.19900 + (-0.00157) + 0.00000 + (-0.00000) + 0.00000 + \cdots] = 0.3949.$$

This *p*-value is greater than $\alpha = 0.05$, suggesting that the samples from the ground and inventory-2 follow the same frequency proportions.

It would be completely wrong if the calculated *D*=0.26667 (*p*-value=0.3949) was interpreted to mean that there was no significant difference between the two sets of frequency counts for the tree height from the ground and inventory-2. The frequency counts for the tree height from the ground and inventory-2 are totally different (Table 45). It would be completely wrong to mix frequency proportions with frequency counts when using the KS test for inferencing.

Therefore, it was emphasized earlier in Section 3.4 that it was critically important to recognize that the KS test only enables us to evaluate if there is a significant difference between two sets of frequency proportions. It does not evaluate if there is a significant difference between two sets of frequency counts (numbers), nor does it test the vaguely defined "distributions" between two samples unless the "distributions" are defined explicitly as the frequency proportions.

The KS test is nonparametric. It does not assume that the two sets of data are sampled from any defined statistical distributions (e.g., normal distributions). It calculates the maximum difference between the two cumulative distributions from two data sets, and reports a *p*-value from that and the sample sizes. The null hypothesis of the KS test is that both data sets (groups) were sampled from populations with identical frequency distributions or cumulative frequency distributions. So long as the frequency distributions or cumulative frequency distributions are statistically identical, the KS test is invariant to different locations (e.g., means, medians) and different variations (e.g., variances, standard deviations).

The Kolmogorov-Smirnov Test for Normality

Sometimes, the two sample Kolmogorov-Smirnov test discussed in this study (which compares two groups of data) may be confused with the one sample Kolmogorov-Smirnov test. Together with the Shapiro-Wilk test, the Cramer-von Mises test, and the Anderson-Darling test, the one sample KS test is most frequently used as a test of normality for one set of data (SAS Institute Inc. 2011). For instance, it has been used to test whether a set of data (e.g., a set of errors) comes from a normal distribution (Huang 2002, Yang et al. 2004, Huang et al. 2019).

Since we are comparing two groups of data (e.g., from the ground and inventory), the two sample KS test is used throughout this study. Readers who are interested in using the one sample KS test to test the normality of a data set may follow the examples described in the above-mentioned references.

Variables Characterizing a Statistical Distribution

Although most people understand the difference between "frequency distribution" and "cumulative distribution", the distinction between "distribution" and "frequency distribution" can sometimes be blurry. Many people consider their difference to be "semantics" and refer to them interchangeably. This may be fine in some day-to-day conversations. But in the context of statistical tests, they must be clearly and explicitly defined. Failing to do so can lead to confusion and misleading conclusions.

As mentioned before (in Section 3.4), in statistics and data science, a "distribution" is typically characterized by three variables: its location, dispersion and shape. Any significant difference between two distributions can be caused by the difference in any one of these three variables. In practice we often use, at a minimum, mean and standard deviation to describe (the location and dispersion of) a statistical distribution. In some more technical contexts, we also include the skewness (a measure of data symmetry) or kurtosis (a measure of data heavy or light tail-ness relative to a normal distribution) to describe the shape of a distribution. For a set of samples (e.g., y_i's) that consists of *n* data points (*i*=1, 2, ⋯, *n*), the mean ($\bar{y}$), standard deviation (SD) and skewness (or kurtosis) are defined by:

[5.38] $$\bar{y} = \sum y_i / n$$

[5.39] $$\text{SD} = \sqrt{\sum (y_i - \bar{y})^2/(n-1)}$$

[5.40] $$\text{Skewness} = \frac{\sum (y_i-\bar{y})^3/n}{\left(\sqrt{\sum (y_i-\bar{y})^2/n}\right)^3} \quad \left(\text{or kurtosis} = \frac{\sum (y_i-\bar{y})^4/n}{\left(\sqrt{\sum (y_i-\bar{y})^2/n}\right)^4}\right)$$

If a statistical test is used to claim that there is no significant difference between two distributions, it must prove simultaneously, that at least two of the three variables from the two distributions, location (mean) and dispersion (variance or standard deviation), are statistically equivalent. Unfortunately, to date, we are unaware of the existence of any single test that can prove this simultaneously, even though there are many tests that can test each individual variable and any significant difference in any one of the individual variables can be used to claim that the distributions are different. In our view, it is a common mistake in the literature to claim that the distributions of the two samples are the same based on the chi-square test or the KS test. This is demonstrated in the above examples, where entirely different distributions in terms of the means and standard deviations led to the same chi-square test or KS test result.

A Further Note on the Kolmogorov-Smirnov Test Applied to Grouped Data

For the KS test, although on the surface it is used to evaluate the difference between two frequency proportions, inherently and more specifically, it is implemented through testing the cumulative proportions. This should be clear from [4.10], and from the graphs in Figures 6, 9, 12 and 16(c) or 16(d).

Sometimes, a continuous variable is divided into user-defined class-intervals (groups). Histograms (or frequency bars, frequency charts) are then drawn to display and compare the frequency distributions of the grouped data. For instance, for Data-2 in Table 16, tree heights from the ground (y) and inventory (x) can be grouped by a 1, 2, 3 m or any other user-defined height class. Table 46 and Table 47 list the data grouped by 2 and 3 m height classes, respectively, alongside the frequency counts that correspond to the height classes.

TABLE 46. TREE HEIGHT COUNTS BY 2 M HEIGHT CLASSES FOR DATA-2 IN TABLE 16.

Type	Tree height												Total
	14	16	18	20	22	24	26	28	30	32	34	36	
Ground frequency count	0	0	3	8	3	3	8	4	8	3	0	2	42
Inventory frequency count	1	1	4	9	6	9	6	3	3	0	0	0	42

Note: 2 m height class ranges are defined by 13≤height<15, 15≤height<17, ⋯, 35≤height<37.

TABLE 47. TREE HEIGHT COUNTS BY 3 M HEIGHT CLASSES FOR DATA-2 IN TABLE 16.

Type	Tree height								Total
	16	19	22	25	28	31	34	37	
Ground frequency count	0	8	7	9	8	7	2	1	42
Inventory frequency count	3	11	9	13	5	1	0	0	42

Note: 3 m height class ranges are defined by 14.5≤height<17.5, 17.5≤height<20.5, ⋯, 35.5≤height<38.5.

Figure 17 shows the histograms of the grouped data by 1, 2, 3 and 4 m height classes. While histograms provide an intuitive and convenient graphical technique for showing the grouped data, their use and interpretation must be done carefully (to avoid becoming part of misleading data visualization).

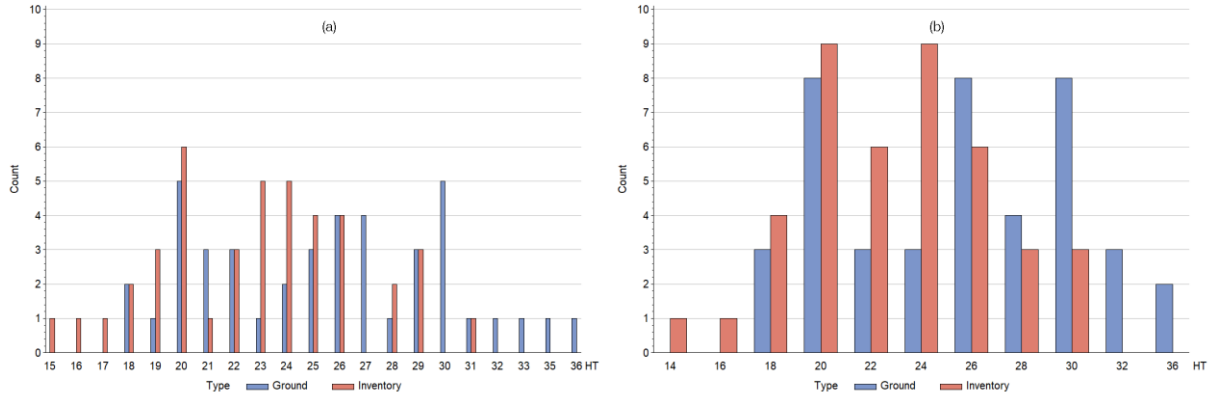

Figure 17 (part 1 of 2). Frequency counts for Data-2 in Table 16 grouped by 1 m (a), 2 m (b), 3 m (c) and 4 m (d) height classes. Grouped data for graphs (b) and (c) are listed in Table 46 and Table 47, respectively.

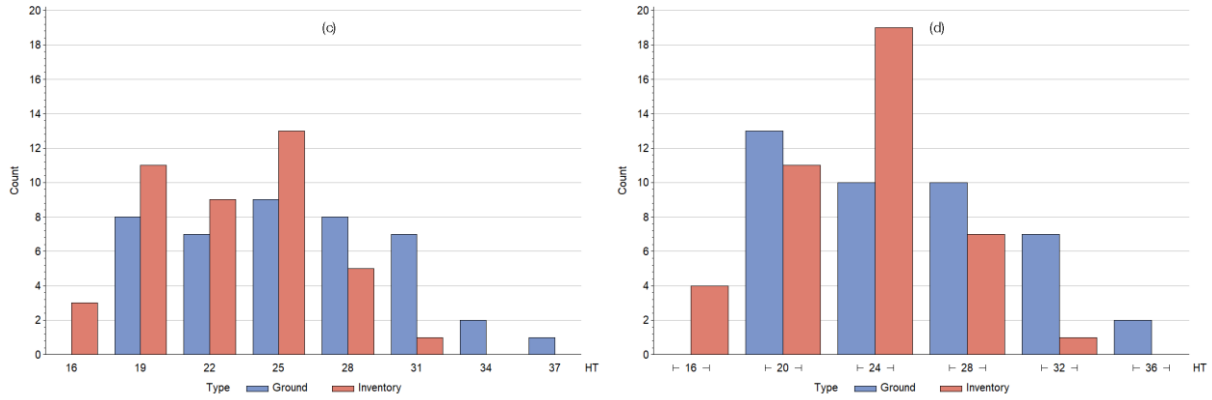

Figure 17 (part 2 of 2. Frequency counts for Data-2 in Table 16 grouped by 1 m (a), 2 m (b), 3 m (c) and 4 m (d) height classes. Grouped data for graphs (b) and (c) are listed in Table 46 and Table 47, respectively.

For instance, the histograms in Figure 17 look quite different or "flexible" depending on the height class intervals we happened to choose. This flexibility, caused by the arbitrary choice of a 1, 2, 3 or 4 m height class (or any other user-defined height class), can lead to misuses by some data analysts, who may be tempted to "pick and choose" or "manipulate" a class interval that satisfies his/her intent. The problem may be exacerbated if the KS test is applied to make inferences and prove the intent.

To illustrate, following the procedures described in Section 4.2, the KS test is applied to Data-2 in Table 16 grouped by 1, 2, 3 and 4 m height classes shown in Figure 17. The calculated *D* statistics and *p*-values that correspond to different class intervals are listed in Table 48. For comparison, the *D* statistic and *p*-value obtained earlier in Section 4.2 for the original (ungrouped) Data-2 are also listed in Table 48.

It is clear from Table 48 that, using the original data, the KS test (*p*-value=0.0188) indicates that the tree heights from the ground and inventory do not follow the same frequency distribution (i.e., frequency proportion). However, when the data are grouped by a 1 m height class, the KS test (*p*-value=0.0649) suggests that the tree heights from the ground and inventory follow the same frequency distribution. But when the data are grouped by a 2 m height class, the KS test (*p*-value=0.0358) returns to indicate that the tree heights do not follow the same frequency distribution. Furthermore, when the data are grouped by a 3 m height class (*p*-value=0.0649) and a 4 m height class (*p*-value=0.1121), the KS tests come back to suggest again that the tree heights follow the same frequency distribution.

TABLE 48. THE KOLMOGOROV-SMIRNOV TEST STATISTICS (AND THEIR *P*-VALUES) FOR ORIGINAL AND GROUPED DATA.

Data	Original (ungrouped)	Tree height grouped by			
		1 m class	2 m class	3 m class	4 m class
Data-2 in Table 16	*D* = 0.3333 (*p* = 0.0188)	*D* = 0.2857 (*p* = 0.0649)	*D* = 0.3095 (*p* = 0.0358)	*D* = 0.2857 (*p* = 0.0649)	*D* = 0.2619 (*p* = 0.1121)

The back-and-forth of the significance in Table 48 is clearly an artifact of the selected class-intervals. Basically it says, depending on the arbitrary choice of the height class intervals, we can get statistically significant and insignificant results for the frequency distributions of tree heights from the ground and inventory, even though the original data are unchanged.

The above example and discussion on grouping a continuous variable into some user-defined class-intervals are very relevant in our subject areas (i.e., forest inventory, forest mensuration and forest modeling), as we frequently group our data into different classes or strata, then analyze them. There are several important implications about grouping, displaying and analyzing such data:

1. Methodologically, while using histograms (e.g., Figure 17) can provide an intuitive and convenient graphical tool for showing the grouped data, they can be "manipulated" to look differently to serve an analyst's intent, by purposely (and arbitrarily) selecting class intervals that satisfy his/her intent. Therefore, their true value and effectiveness in analyzing continuous data are limited and questionable, albeit that they can be quite useful for categorical data. We should not overuse or rely on them for continuous data. They may not really tell much, and they are prone to bias caused by the arbitrary choice of the width and location of the chosen class-intervals. The shape of the histograms can always be changed or "manipulated" to look better or worse depending on the arbitrary choice of class-intervals. Intentional or deliberate choice of class-intervals can make asymmetrical histograms look more symmetrical, and vice versa.

2. When a continuous variable is grouped into some user-defined class-intervals, any statistics and statistical tests based on the class-intervals are prone to bias, because it can always be argued that the class-interval boundaries are inherently arbitrary, and that some alternative class-intervals can yield very different results. These are clearly explained in Dransfield and Brightwell (2012) and illustrated in Table 48. Arbitrary elements and choices in statistical analysis are always potentially biased. They should be avoided whenever possible. The process of grouping a continuous variable into different class-intervals always results in the loss of more detailed information.

3. In many practical applications, we often group, e.g., tree DBH and tree height into some pre-determined class-intervals, such as 2 cm or 1 inch DBH classes, and 1, 2, 3, 5 or 6 m height classes. It is imperative to recognize that any statistics and statistical tests obtained from the specific class-intervals only pertain to those class-intervals used in the analysis. They must be interpreted and used as such in inferencing and applications. They cannot be used to make any broad generic inferences, or changed in the middle of inferencing and applications.

4. Unless a particular objective or constraint dictates otherwise, it is always preferable and better to analyze a continuous variable in its original undivided form, and to conduct statistical analysis on original, rather than grouped, normalized or transformed data. Results and inferences derived from the grouped, normalized or transformed data only apply to such data, not to the original data. Practitioners should be aware that, no matter how good the results may appear on grouped, normalized or transformed data, they may not be relevant to the original data.

One other important implication worth highlighting from the above example in Table 48 is that, statistical significance can be changed back-and-forth depending on the arbitrary choice of class-intervals. It shows how easy it can be, to potentially "pick and choose" and misuse "statistical significance" in inferencing and decision-making. We provide some of our critical thoughts on this next, while fully recognizing that there is no "best solution" for this and that different viewpoints, objectives and philosophies exist.

5.5 Statistical significance and practical significance

It is really unfortunate that "statistical significance chasing" or "statistical significance fishing" has been entrenched into some scientific fields, resulting in many questionable, misleading and erroneous conclusions (and causing the "damning" of statistics by some researchers and practitioners alike). This is largely caused, intentionally or unintentionally, by the misunderstanding and misuse of statistical and data science, rather than by the science itself. It has led to some true experts in their fields to caution that "most of what is published in journals is just plain wrong or nonsense" (Richard Smith, former BMJ (British Medical Journal) editor-in-chief, quoted in Freedman 2010), and that "most published research findings are false and untrustworthy" (Ioannidis 2005, Smith 2014, Lose and Klarskov 2017).

If not careful, statistics and particularly statistical significance/insignificance can add the air of scientific rigor to bad research and help researchers fool themselves and research users. It is often easier to get a peer-reviewed paper published if one uses erroneous statistical analysis than if one uses no statistical analysis at all (Hurlbert and Lombardi 2003). Chalmers and Glasziou (2009) and Glasziou and Chalmers (2016) noted that a large percentage (85%) of research in biomedical fields is "wasted", echoing what was already observed by Altman (1994) more than a quarter century ago that, "huge sums of money are spent annually on research that is seriously flawed through the use of inappropriate designs, unrepresentative samples, small samples, incorrect methods of analysis and faulty interpretation", and that "we need less research, better research and research done for the right reasons". Feyerabend (1981) lamented that most researchers today are devoid of ideas, intent on producing some paltry result "so that they can contribute to the flood of inane number of papers that now constitutes 'scientific progress' in many areas". As pointed out forcefully by Ioannidis (2005) and reaffirmed by Smith (2014), in "many current scientific fields, claimed research findings may often be simply accurate measures of the prevailing bias", and "most scientific studies are wrong, and they are wrong because scientists are interested in funding and careers rather than truth".
While it is understandable that for various reasons some researchers may not like or agree with the criticisms raised by the above experts, many researchers do recognize that a large percentage of the published papers in biology contain serious statistical mistakes. This is well-documented in Dransfield and Brightwell (2012) and Makin and Orban de Xivry (2019). We should at least be aware of the potential traps and pitfalls of statistical significance and avoid the common statistical mistakes observed by many experts. More importantly, we need to realize that statistical significance (or any single statistic or test) is just a part of many factors that may need to be considered during inferencing and decision-making. It is not the sole deciding factor. In particular, forest practitioners should recognize the difference between statistical significance (i.e., if the *p*-value is less than a specified significance level, typically at α=0.05) and practical significance (i.e., if the magnitude of the difference is large enough to have any important or meaningful practical consequence), and be able to use them appropriately during inferencing and decision-making.

Statistical significance defined by *p*-values is impacted directly by sample size and sample variability. It might seem logical that statistical significance and *p*-values relate to importance and causality. But this is not true. With enough samples or low

sample variability, even a tiny difference will result in a statistic or a statistical test to be statistically significant. But statistical significance does not imply that the difference has any practical real world consequence. It is not necessarily related to importance, nor to causation. The tiny difference may be trivial, inconsequential and meaningless in practice even through it is statistically significant.

Practical significance is not directly impacted by sample size and sample variability. While statistical significance tells whether the difference exists, practical significance tells whether the magnitude of the difference is practically significant. What is practically significant and meaningful may sound subjective and can depend on the specific situation and objectives, but it is typically determined by trained professionals based on the specialized knowledge, expertise and experience on the subject. These knowledge, expertise and experience include relevant biological and operational considerations, the inherent variation of the variable and data in interest, and other subject matter understanding and goals, as well as the operational impact of the difference or "error".

We described several statistical tests in this study. They can be used to answer some specific questions from the repeated sampling point of view. They can also help to guide the direction of further analysis. The practical importance of the test results, however, is judged differently, as it is highly impacted by many other factors. In addition, the power of these tests is highly influenced by the sample sizes. A small sample size may result in insufficient power to detect a true difference between two techniques. A large sample size may result in very high power, making practically trivial and non-meaningful differences appear statistically significant. In fact, almost all statistical tests will result in the rejection of the null hypothesis if the sample size is very large. Aronoff (1982), Fitzgerald and Lees (1994), Stehman (1997), Fleiss et al. (2003), Foody (2004), Dransfield and Brightwell (2012) and Stehman and Foody (2019) provided good examples and discussions on this topic. But a practical dilemma is that, how "small" is too small and how "large" is too large are generally not quantitatively defined. They depend on many factors, and there seems to be no indication from the literature we have found as to what might constitute a "suitable subjective balance". Therefore, we suggest that the test results, wherever relevant, be considered only as approximations and as one of the factors in holistically judging forest inventory techniques. Calculations of the descriptive statistics and inferential statistics for categorical and continuous variables are just one of the steps in any rigorous data analysis and inference.

It is always greatly preferable and often necessary, to jointly consider statistical significance together with practical significance, which, depending on the subject areas involved, has also been referred to as clinical significance, biological/*ecological* significance, *economical significance,* or mensurational significance (e.g., Bland and Altman 1986, Engsted 2009, Schober et al. 2018, Huang et al. 2019). Practical significance is often more important and meaningful than statistical significance in many situations. In any analysis we should always avoid statistical significance chasing and "data dredging" (or *p*-hacking, data snooping, data butchery, data torturing, data massaging, data nursing). Interested readers are strongly recommended to read and comprehend the statements by the American Statistical Association (Wasserstein and Lazar 2016), and the articles by Ioannidis (2005), Lambdin (2012), Makin and Orban de Xivry (2019), and especially Nuzzo (2014).

5.6 Calibration and localization for remotely sensed inventory data

In practice, it is unlikely that two sets of measurements for the same variable, one obtained on the ground and the other from an inventory technique, are in perfect or near perfect agreement. Random and/or systematic errors almost always occur in any type of measurement. Random errors are part of the natural variation. They are more difficult to remove or reduce. However, systematic errors in the measurements (classifications, predictions) can always be removed or calibrated to the "observed truth" on the ground, or to a commonly accepted consensus or gold (reference) standard.

Figure 18 illustrates two imperfect scenarios in which the ground measures (y_i or simply y) differ from the corresponding inventory measures/estimates (x_i or simply x). The differences between the measures can be positive or negative, linear or nonlinear in nature. Assuming that the ground measures are the truth or the accepted reference (gold) standard, inventory estimates often need to be adjusted or calibrated to the ground measures, so when the situation warrants, the purportedly more effective and efficient (i.e., reduced costs at increased scope and speed of data collection) inventory estimates may be used to substitute the ground measures (more common), and vice versa (less common).

There are different ways to calibrate a set of measurements against a different set of measurements. They belong to the general concept of method (or model) calibration, sometimes also referred to as method adjustment, method modification, correction, conversion, or localization (Huang 2002, Huang et al. 2019). If an analysis or an agreement study shows that two measurements, one from the ground and the other from an inventory are the same and thus interchangeable, there is no need to do any method calibration. However, if the analysis shows that the two measurements are not the same, two follow-up actions may be taken.

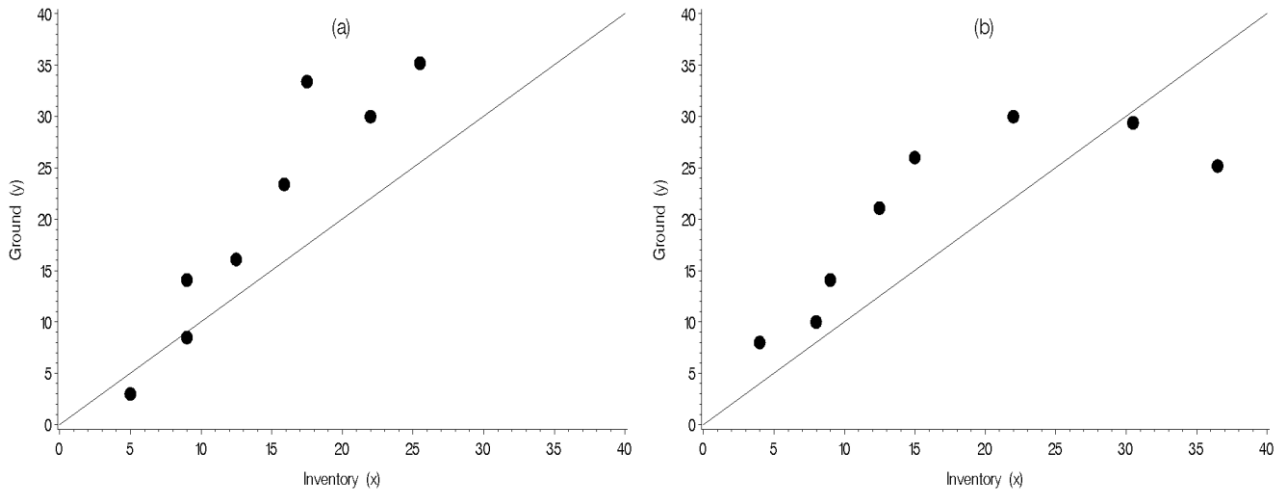

Figure 18. An illustration of imperfect agreement scenarios between the ground (*y*) and inventory (*x*) measures, where the diagonal line is the 45° line that passes through the origin. It represents perfect agreement (*y*=*x*). The imperfections can be positive or negative, linear or nonlinear in nature.

In the first case, one may want to calibrate a set of measurements from a new inventory method against another set of measurements from the ground or a well-established method that is considered the truth or the accepted reference standard, so that the new method, which supposedly represents a simpler or a more effective and efficient method, is in-line with the truth or the accepted reference standard. In the second case, one may not know the truth or the accepted reference standard, but just want to establish an intrinsic quantitative functional relationship between the two sets of measurements, such that whenever the situation presents or requires, one set of measurements can be converted into the other, and vice versa. In both cases the concept of method calibration comes up.

Regardless of the cases, in broad mathematical and statistical terms, the concept of method calibration by and large involves the fitting of one of the following regression models expressed in the general forms of:

[5.41] $y_i = f(x_i)$

[5.42] $y_i = f(x_i, \text{other potential variables})$

where y_i denotes the ground (or reference) measure and x_i denotes the corresponding inventory measure (or classification, estimate, prediction) for the ith observation (i=1, 2, ⋯, n, n is the total number of observations), f denotes a linear or nonlinear function, and "other potential variables" denote other potential inventory (more common) and ground (less common) variables/measures/metrics that may contribute to or can explain the difference between y_i and x_i.

For many practical applications, [5.41] expressed in a simple linear form (i.e., $y_i = b_0+b_1x_i$, where b_0 is the intercept and b_1 is the slope) may be all that is needed. When "other potential variables" are included in the regression in [5.42], the sources of the discrepancies between ground (y_i) and inventory (x_i), and the exact amount of the impact of the "other potential variables" on calibration, can be quantified in addition to that of x_i. Readers who are interested in more detailed descriptions about the methods for calibration and localization can look into Huang (2002), Yang and Huang (2014) and Huang et al. (2019), where step-by-step examples are provided.

Here, only the most commonly used calibration method based on the ordinary least squares (OLS) fit of the simple linear regression is shown. The felled tree height data (Table 20) and the stand density data (Table 24) from the ground and the lidar inventory in the FMA area of Canadian Forest Products Ltd. (Grande Prairie) are used again to demonstrate the calibration method.

Calibrating Lidar Height to Felled Tree Height

In Section 4.3, it was shown that based on the results in Figures 7-8, Table 21 and the KS test, it can be inferred that the agreement between felled height and lidar height is reasonably good. Lidar height and felled height can be used interchangeably. However, the results shown in Table 21 indicate that overall, there still is a $\bar{e}$=-0.371 (m) or $\bar{e}$%=-2.2% prediction bias from lidar. If one wants to achieve a bias-free prediction on average from lidar (relative to the felled heights), the following simple linear regression can be fitted:

[5.43] $HT_{felled} = b_0 + b_1 \times HT_{lidar}$

where the felled height (y_i, HT_{felled}) and lidar height (x_i, HT_{lidar}) are given in Table 20, b_0 is the intercept and b_1 is the slope, estimated to be b_0=0.01039 and b_1=0.97787 from the OLS fit. The root mean square error of the fit is $RMSE_r$ = 1.543 (from [4.21]), and the coefficient of determination is R^2 = 0.912 (from [4.22]), which describes the correlation between HT_{felled} and HT_{lidar}. Figure 19(a) shows the fit (the dashed line).

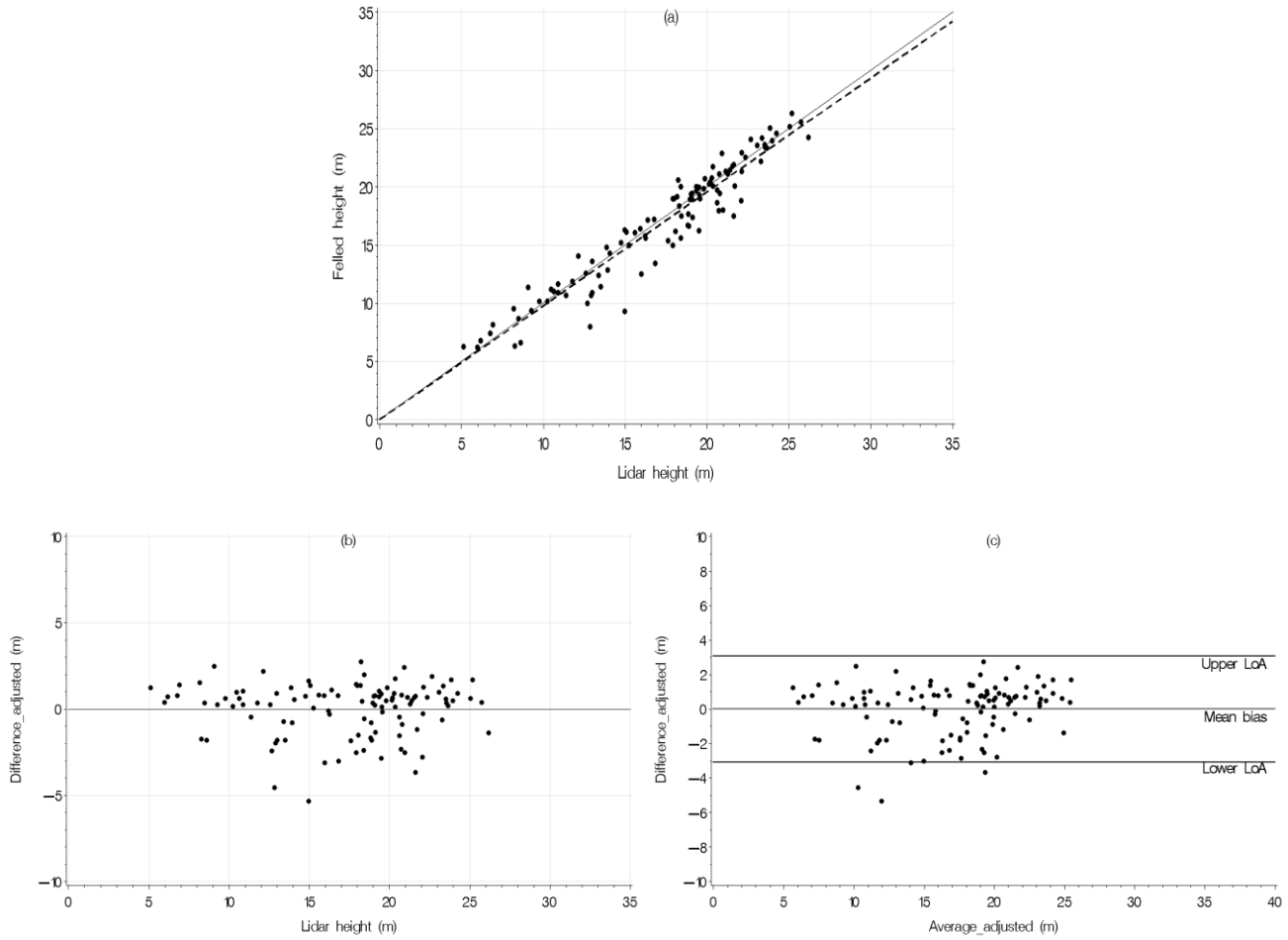

Figure 19. Scatter plot (a), error plot (b) and the Bland-Altman plot for felled heights and lidar heights after the adjustment. Actual data are listed in Table 20. The solid line in (a) is the 45° line. The dashed line in (a) is the fitted regression from [5.43]. The LoA lines in (c) are limits of agreement.

The fitted regression line defined by [5.43] can be used to adjust the lidar heights (i.e., to predict felled heights from lidar heights). The errors after the adjustment (e_{adj}) are the residuals from the OLS fit, calculated by:

[5.44] $e_{adj} = HT_{felled} - \widehat{HT}_{felled} = HT_{felled} - (b_0 + b_1 \times HT_{lidar})$.

Graph (b) in Figure 19 shows the errors after the adjustment (difference_adjusted) on the *y*-axis, plotted against the lidar heights on the *x*-axis (standard residual plot of residuals against the predicted felled heights from the OLS fit is available but not shown here). Graph (c) in Figure 19 shows the Bland-Altman plot after the adjustment. Since the mean of the adjusted errors each calculated by [5.44] is zero, an OLS property (Draper and Smith 1998), after the adjustment the predictions of felled heights from lidar heights are bias-free on average. This is shown in Table 49 (first row), along with other goodness-of-fit statistics and agreement measure calculated after the adjustment.

TABLE 49. GOODNESS-OF-FIT STATISTICS AND AGREEMENT MEASURE AFTER THE ADJUSTMENT.

Type	*n*	$\bar{e}$	MAE	RMSE	$\bar{e}$%	MAE%	RMSE%	e_{10}	e_{33}	e_{50}	MOA
Tree height	108	0	1.205	1.528	0	7.1%	9.1%	0.713	0.981	0.981	0.954
Stand density	28	0	193.6	243.5	0	27.7%	34.8%	0.214	0.679	0.750	0.685

Note: tree height data are listed in Table 20, stand density data are listed in Table 24, *n* denotes the sample size, $\bar{e}$, MAE, RMSE, $\bar{e}$%, MAE%, RMSE%, e_{10}, e_{33}, e_{50} and MOA are defined in [4.1]-[4.5] and [4.7].

Compared to the results from the unadjusted goodness-of-fit statistics and agreement measure in Table 21 and the graphs in Figures 7-8, the gain from the adjustment is not obvious. This is expected because for this example, the agreement between the original (unadjusted) lidar heights and felled heights is already reasonably good.

The comparison also reveals that, although after the adjustment the predictions of felled heights from lidar heights are bias-free on average (i.e., $\bar{e}$=0 and $\bar{e}$%=0, Table 49, first row), other goodness-of-fit statistics are not guaranteed to be better after the adjustment. For instance, after the adjustment MAE=1.205 (Table 49), which is larger than that before the adjustment (MAE=1.128, Table 21). This is because that on average, the adjustment will achieve bias-free (the individual errors sum up to zero), but the size of each individual error after the adjustment is not guaranteed to be smaller. It is the case of average versus individual errors. An improved average does not necessarily mean an improvement for all individuals.

Notice also that the RMSE in Table 49 (RMSE=1.528) is different from the $RMSE_r$ from the regression fit ($RMSE_r$=1.543), notwithstanding that the difference is small in this case. But it is important to understand the exact mathematical difference between RMSE and $RMSE_r$.

Calibrating Lidar Density to Ground Density

For the stand density data in Table 24, the following simple linear regression can be fitted:

[5.45] $N_{ground} = b_0 + b_1 \times N_{lidar}$

where N_{ground} is the ground density (stems/ha) and N_{lidar} is the lidar density (stems/ha).

The intercept and the slope for [5.45] are estimated to be: b_0=71.20421 and b_1=1.39811. The root mean square error and the coefficient of determination from the OLS fit are $RMSE_r$=252.7 and R^2=0.521, respectively. Figure 20(a) shows the fit (the dashed line).

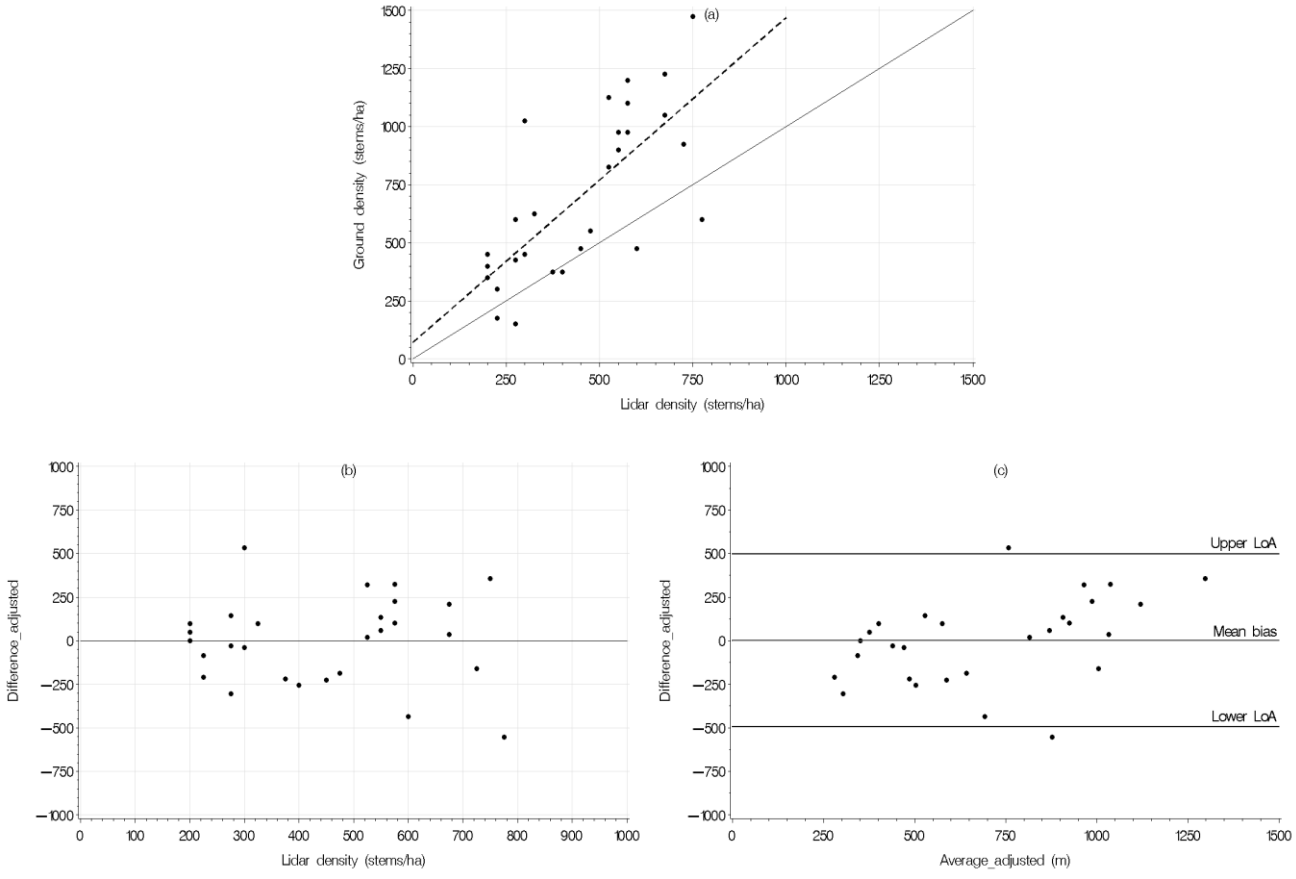

Figure 20. Scatter plot (a), error plot (b) and the Bland-Altman plot for ground densities and lidar densities after the adjustment. Actual data are listed in Table 24. The solid line in (a) is the 45° line. The dashed line in (a) is the regression from [5.45]. The LoA lines in (c) are limits of agreement.

The fitted line defined by [5.45] is used to adjust the lidar density. The errors after the adjustment (e_{adj}) are the residuals from the OLS fit, calculated by:

[5.46] $e_{adj} = N_{ground} - \hat{N}_{ground} = N_{ground} - (b_0 + b_1 \times N_{lidar})$.

Graph (b) in Figure 20 shows the errors after the adjustment (standard residual plot of residuals against the predicted ground densities from the OLS fit is available but not shown here). Graph (c) in Figure 20 shows the Bland-Altman plot after the adjustment. Once again, after the adjustment the predictions of ground densities from lidar densities are bias-free on average (i.e., $\bar{e}$=0 and $\bar{e}$%=0). This is also shown in Table 49 (second row), along with other goodness-of-fit statistics and agreement measure calculated after the adjustment.

Compared to the results from the unadjusted goodness-of-fit statistics and agreement measure in Table 25 and the graphs in Figures 10-11, the gains from the adjustment are obvious. For instance, after the adjustment the RMSE is reduced from 356.4 (Table 25) to 243.5 (Table 49, second row), the absolute error (MAE) is reduced from 285.7 to 193.6, the measure of agreement MOA is increased from 0.421 to 0.685, and the e_{10} is increased from 0.107 to 0.214. The adjusted errors are scattered (more or less) homogeneously in a band around the zero line (Figure 20(b)), whereas the unadjusted errors are clearly skewed upwards above the zero line (Figure 10(b)), implying that if unadjusted, the lidar densities would underestimate the ground densities (by an average of 250 stems/ha, Table 25).

However, when comparing the adjusted (Figure 20(c)) and unadjusted (Figure 11(a)) Bland-Altman plots for ground densities and lidar densities, the number of data points outside the LoA lines in the Bland-Altman plot after the adjustment increased from zero (Figure 11(a)) to two (Figure 20(c)). This is due to what was mentioned earlier, that on average the adjustment will achieve bias-free (the individual errors sum up to zero), but the size of each individual error after the adjustment is not guaranteed to be smaller. The overall improvement of Figure 20(c) over Figure 11(a), though, is obvious. Whether the improvement is good enough is a different question, as after the adjustment, there are 26 out of 28 observations (26/28 =93%) fall within the LoA lines. This is slightly lower than the 95% specified by Bland and Altman (1986) for good agreement.

Further improvement to the stand density calibration can be explored by incorporating additional variables on the right hand side of the calibration model, or by dividing or grouping the data into more homogenous subgroups or strata (e.g., based on height, density and/or geographical area) and fit the models by the strata. For instance, [5.45] may be expanded or changed to:

[5.47] $N_{ground} = b_0 + b_1 \times N_{lidar} + b_2 \times H_{DC}$

[5.48] $N_{ground} = b_0 + b_1 \times N_{lidar} + b_2 \times H_{top}$

[5.49] $N_{ground} = b_0 + b_1 \times N_{lidar} + b_2 \times H\% + b_3 \times CC$

[5.50] $N_{ground} = b_0 + b_1 \times N_{lidar} + b_2 \times H_{DC} + b_3 \times H_{top} + b_4 \times CC$

[5.51] $N_{ground} = b_0 + b_1 \times N_{lidar} + b_2 \times H_{DC} + b_3 \times H_{top} + b_4 \times H_{std} + b_5 \times CC$

[5.52] $N_{ground} = b_0 N_{lidar}^{b_1} H_{DC}^{b_2}$

[5.53] $N_{ground} = b_0 N_{lidar}^{b_1} H_{DC}^{b_2} \exp(H_{top}^{-b_3})$

[5.54] $N_{ground} = b_0 N_{lidar}^{b_1} H_{DC}^{b_2} \exp(H_{top}^{-b_3})\ CC^{b_4}$

where all the right hand side variables are extracted from lidar, H_{DC} is the dominant and co-dominant height, H_{top} is the top height, H% is the top percent height (see Section 4.4 and Table 23), CC is the crown closure (either in percentage or in actual size/area), and H_{std} is the standard deviation (std) of lidar extracted tree heights.

Numerous additional lidar-derived variables based on height, crown cover and vertical structure metrics (or even some variables obtained on the ground) can also be evaluated and incorporated into [5.47]-[5.54], provided that they are readily available or easy to obtain and can explain additional variations (Næsset et al. 2005, Li et al. 2008, Bater et al. 2011). However, the minimum (H_{min}) and maximum (H_{max}) of lidar extracted tree heights are often not used in the models, because H_{min} is almost always the same number (whatever minimum threshold chosen) and H_{max} might be an outlier or a highly influential data point (which may invoke the use of robust regression – we will not go there in this study). Instead, height percentiles such as the 25th and 95th percentiles are typically used (Bater et al. 2011). Other functional forms can also be explored. They all fall within the general expression given by [5.42].

The Dependent Variable and Independent Variables in Calibration

There are two ways to develop a regression function between ground measure and inventory measure, depending on whether the ground measure or the inventory measure is used as the dependent variable (the *y*-variable):

[5.55] Ground measure = *f*(inventory measure)

[5.56] Inventory measure = *f*(ground measure)

where ground measure denotes a ground-measured forest attribute, inventory measure denotes the forest attribute extracted or derived from an inventory technique, and *f* denotes a general expression for a linear or nonlinear function. For instance, corresponding to the above general expressions, for ground height and lidar height, we could have:

[5.57] Ground height = $a_1+b_1\times$lidar height.

[5.58] Lidar height = $a_2+b_2\times$ground height.

If the standard OLS method is applied to estimate the parameters in [5.57] and [5.58], we would get two distinct lines for the same set of data, because the OLS estimates of both the intercepts (a_1 and a_2) and the slopes (b_1 and b_2) would be different. This would impact calibration, as two inconsistent ground values could be predicted, one directly from [5.57], and the other indirectly from [5.58] through inversing, i.e., through re-arranging [5.58] and solving for ground measure (ground height in this example, producing ground height = (lidar height - a_2)/b_2).

In remote sensing studies, the choice of which one to use between [5.55] and [5.56] appears quite arbitrary. Means et al. (1999), Heurich et al. (2004), Zhao et al. (2018) and Liu et al. (2018) chose to use [5.55] while Lim et al. (2001, as cited in St-Onge et al. 2003), Coops et al. (2007), Kwak et al. (2007), Sibona et al. (2017), Wang et al. (2019), Krause et al. (2019) and most of the others chose to use [5.56]. While the choice of which variable to use as dependent or independent variable is inconsequential in a simultaneous equation system where a variable appearing on the right-hand side of an equation can also appear on the left-hand side of another equation in the system (e.g., Huang and Titus 1999), most of the equations developed in remote sensing studies are separate individual equations. Therefore, which variable to use as dependent or independent variable is very consequential (Piñeiro et al. 2008).

In addition, since in practice the predominant reason for developing a regression function is almost always to predict a more difficult and costly variable from a relatively simpler and less expensive variable (or several relatively simpler and less expensive variables), it is the optimum to fit the regression function using the more difficult and costly variable as the *y*-variable, because the OLS method minimizes the errors of the *y*-variable, not the *x*-variable(s). Therefore, it is always better to fit a regression function using the ground measure as the *y*-variable if the main purpose is to predict the ground attribute from remote sensing metrics, or to calibrate the remote sensing estimate to the ground measure. In the ground height-lidar height example in [5.57]-[5.58], since ground height is usually more difficult and costly to measure, and lidar height may be simpler and less expensive to measure (on a large scale), we would develop Ground height = $a_1+b_1\times$lidar height, not the other way around. Using lidar height as the *y*-variable produces non-optimum results when predicting ground height.

It could always be argued that for theoretical and non-practical academic reasons, one may only be interested in understanding the intrinsic quantitative functional relationship between the two sets of measurements from the ground and inventory. Achieving the best prediction for a more difficult and costly variable from a simpler and less expensive variable may not be the goal. If that is the case, it would be hard to refute, as we cannot prevent people from using a more difficult and costly variable to predict a simpler and less expensive variable. But that would be contrary to common sense.

Furthermore, if indeed one may not know which variable can be considered the truth and used as the *y*-variable, but instead just want to establish an intrinsic quantitative functional relationship between the two sets of measurements, then the OLS method is inadequate, or at least non-optimum. One should implement the orthogonal distance regression, geometric mean regression, OLS-bisector regression or arithmetic mean (OLS-mean) regression. Each of these four regression methods treats both dependent variable (*y*) and independent variables (*x*) equally or symmetrically (Isobe et al. 1990, Babu and Feigelson 1992, Van Huffel 1997). But still, none of them is as good as the OLS method if a goal or a by-product is to obtain the best prediction for a target variable, be it *y* or *x*. Since our goal of developing a regression function is almost always to predict a more difficult and costly variable from a simpler and less expensive variable (or variables when relevant), we will not discuss the symmetric regression methods further in this study. Interested readers may want to look into Huang et al. (2019) on the technical details of these four symmetric regression methods and their comparison to other methods.

Approximation through Proportional or Ratio Adjustment

A simple proportional or ratio adjustment method can also be used for calibration. This method is implemented by calculating the proportion (P) or ratio between the average of the predictions from an inventory ($\bar{x}$) and the average of the corresponding ground observations ($\bar{y}$):

[5.59] $P = \frac{\bar{x}}{\bar{y}}$

Once the P is available, the values of any existing or future ground observations are calibrated as:

[5.60] Ground value (y) = $\text{prediction } (x)/\text{P}$

For instance, if on average lidar heights (derived from a lidar metric or metrics) are found to be 90% (P=0.9) of the ground heights, any lidar height can be calibrated to obtain ground height via: ground height=lidar height/0.9.

The proportional adjustment method can be equally as effective as other more complex calibration methods if the data sets from the ground measures and inventory predictions meet certain conditions (e.g., linearity and proportionality between the data sets across the observation and prediction ranges). This method is described in more details elsewhere (e.g., Huang 2002, Huang 2016, Huang et al. 2016). It is quite straightforward and will not be repeated further in this study.

Calibration through Mixed-Effects Methods

Some more complex and computationally more demanding methods for calibration can be achieved through mixed-effects methods. Interested readers may want to start with the step-by-step examples first (Huang 2016, Huang et al. 2016), then, if necessary, move on to incorporate correlated error structures (Meng and Huang 2009, 2010; Meng et al. 2012), and ultimately, generalized error structures (Huang et al. 2009a, Yang and Huang 2011a, Huang et al. 2011). Detailed descriptions on the mixed-effects methods, along with the conditions and input required for using these methods in practice, are presented extensively elsewhere (e.g., Huang et al. 2009b, 2009c; Meng and Huang 2009, Yang and Huang 2011b). They will not be elaborated and repeated further in this study.

A Note of Caution on Calibration

Using a regression function expressed in the general forms of [5.41]-[5.42], or the simple proportional or ratio adjustment method, any measures predicted or extracted from an inventory technique, no matter how bad or good they are, can always be calibrated to the ground measures bias-free (i.e., $\bar{e}$=0 and $\bar{e}$%=0).

However, regression (correlation, association) analysis is very different from accuracy assessment and agreement analysis. Whether the calibrated measures are accurate enough or are in agreement with the ground measures is a different question. The answer to it depends on other factors, including the dispersion or precision of the calibrated measures (e.g., in terms of RMSE, e_{10}, e_{33} and e_{50}), the agreement between the calibrated measures and the ground measures (e.g., in terms of MOA, the error plot and the Bland-Altman plot), and the subject matter biological, mensurational and operational considerations for the specific variable in interest and the objectives of a study (besides the needs to consider time, relevant expertise and requirements, cost and economic viability and other potential variables and constraints). Calibration is not a panacea nor an alchemy for any inaccurate and inoperable measures from inventory techniques. It is always the best to aim for accurate estimates in the first place without the thought of calibration, which may be considered the last resort, or sometimes, the last "salvage" operation.

6 Conclusions and recommendations

The conclusions and recommendations reached based on the analyses conducted in this study are summarized as follows:

1. To be considered an operationally valid forest inventory technique or a preferred forest inventory technique among the competing inventory techniques at the tree and stand level, the inventory technique at a minimum must demonstrate the accuracy and agreement to ground measures in predicting tree species, species composition (species frequency distribution), and species-specific height and density to a meaningful tagging limit. The statistical methodologies presented in this study allow for the determination of the accuracy and agreement for these four primary inventory variables that are critically important to strategic and operational forest management, and that can be extracted or derived from promising new inventory techniques at the tree and stand level.

2. To determine the accuracy of tree species classification/prediction, an error matrix in the form of Table 2 should be provided. Within this error matrix, four accuracy measures should be calculated: the overall accuracy for all species combined (P_o), the correct proportion relative to the ground reference (PR) and the correct proportion relative to the classification (PC) for each species, and the pooled average of the correct proportions for each species (PAve). No other accuracy measures should be computed from the error matrix, as they generally obscure, distract, dilute, confuse or even mislead the real accuracy assessment for tree species classification, and many of them involve questionable formulations or are very idiosyncratic to the specific situation that bears little relevance to the reality of tree species classification.

3. For tree-based approach, prior to assessing the accuracy of species classification in the form of an error matrix, it is important to look at the accuracy of crown delineation or stem segmentation. The overall accuracy of species classification for an individual tree-based inventory is determined by the crown delineation accuracy, as well as the species classification accuracy. It is very important to at least take a look at the proportion of the stems/crowns that are missed by crown delineation. Without assessing and understanding the crown delineation or stem segmentation errors (at least the missing errors), implementation of any tree-based inventory technique in operations must be exercised with great caution and caveats. In any case, without assessing, understanding and finding ways to address the crown delineation or stem segmentation errors, practitioners should not adopt a tree-based approach or any approach as a data collection tool in place of measuring PSPs and TSPs on the ground. Ground-measured PSPs and/or TSPs are necessary to act as the truth or reference standards when determining the accuracy and judging the validity of any forest inventory techniques.

4. The equivalence between species compositions (i.e., species frequency distributions) from the ground measures and inventory measures can be assessed using the chi-square test or Fisher's exact test. In general, the chi-square test should be used. But if more than 20% of the cells in the matrix tabulated for the chi-square test have expected values less than five, then Fisher's exact test should be implemented instead. However, if Fisher's exact test fails to execute due to a large sample size or a large matrix, the chi-square test is a valid approximation and a suitable replacement. The accuracy of the chi-square test increases with the increasing sample sizes and the dimensions of the matrix.

5. When judging the accuracy and agreement (to ground measures) for height and density, at least one of the error plots and three of the goodness-of-fit statistics described in Section 4.1 should be evaluated and included. One of the three goodness-of-fit statistics must be the root mean square error (RMSE) or mean squared error (MSE), which is an overall accuracy measure that combines both bias and precision. For more detailed analysis, especially when focusing on assessing the agreement between ground measures and inventory measures, Mielke's measure of agreement (MOA) and the Bland-Altman plot and analysis should also be included, and the Kolmogorov-Smirnov test should also be conducted.

6. The validity of an inventory technique shall not and cannot be judged by a single matrix, a single statistic or a single test, no matter how powerful it is claimed to be. It is critically important to assess the performance of all four primary inventory variables collectively, so the accuracy and the level of agreement of an entire inventory technique can be determined holistically.

7. Among many potential stand heights, the tree height-ranked top height (H_{top}, the average height of the 100 tallest trees per hectare), the top percent or percentile height (e.g., the average height of the tallest 20% or 25% of the trees per unit area), and the more precisely defined dominant and co-dominant height or overstory height (H_{DC}, the average height of all trees taller than or equal to 80% of the tallest tree per unit area), are recommended for aerial-based forest inventories. The H_{DC} and/or H_{top} are generally preferred in practice, for they are precisely defined and thus repeatable and can be consistently implemented no matter who is using them. Both of them also do not require a full tree list, nor the prediction of tree diameters as a priori.

8. There are a number of technical and operational caveats in assessing forest inventory techniques and applying the statistical methods for the assessment. They include: (1) the three statistical tests used in this study (the chi-square test, Fisher's exact test and the Kolmogorov-Smirnov test), only evaluate the difference between two sets of frequency proportions from two methods, not two sets of frequency numbers, nor two sets of broadly and vaguely defined "distributions"; (2) it is necessary to focus on assessing the key variables that are directly extracted from an inventory technique. Good accuracy and agreement for any other indirectly derived variables such as volume, biomass and carbon, although highly desirable, may not be really indicative of the accuracy and agreement of an inventory technique (it may indicate the quality of ancillary data, the "helper" models or processes involved, the calibration technique chosen, or the viability and strength of the connecting relationships implicated); (3) the standard scatter plot and calibration may mask the true accuracy and agreement of an inventory technique. They need to be used and interpreted with great caution; (4) many conventional idioms and measures associated with the error matrix were derived from questionable concepts and formulations. They have confused many researchers and practitioners alike, and their continued use in assessing the classification accuracy in photogrammetric and remote sensing studies, is unwarranted and should be avoided; and (5) results from any statistical tests should be considered only as approximations and as one of the factors in holistically judging forest inventory techniques. Statistical significance is not necessarily related to importance, nor to causation. It should always be looked at in conjunction with relevant practical, biological and mensurational significance, which can be more important and meaningful than statistical significance in many practical situations.

Comparing new and emerging forest inventory techniques to ground measures can be an intricate and challenging process. The intricacies arise because the comparison typically involves categorical and continuous variables obtained in various qualitative and quantitative forms. The methods presented in this study collectively should provide the right tools capable of addressing and answering the most relevant questions related to such comparisons. There are other considerations in judging a forest inventory technique (such as time, relevant expertise and required resources, cost or economic viability, sampling and other potential operational variables and constraints), which are beyond the scope of what has been discussed in this study. Ultimately, the validity and usefulness of any new and advanced forest inventory technique is determined by how accurate and timely it is in mirroring on-the-ground reality (i.e., the actual forest attributes, characteristics, structures and conditions), and how effective, efficient, reliable, consistent, user-friendly and economically viable it is in real world applications. We are fairly confident that in due course, with the tremendous ongoing efforts by numerous forest inventory specialists, consultants, researchers, academics and other stakeholders, accurate future forest inventories based on Fully Automated Census Techniques (FACTs) at different levels (e.g., tree, stand, landscape and population), will become a reality.

7 References

Agresti, A. 2013. *Categorical data analysis*. 3rd ed., John Wiley & Sons, Hoboken, New Jersey, USA. 752 p.

Agresti, A. 2018. *An introduction to categorical data analysis*. 3rd ed., John Wiley & Sons; Hoboken, New Jersey, USA. 400 p.

Aigner, D.J. 1972. A note on verification of computer simulation models. *Management Science* 18 (11): 615-619.

Altman, D.G. 1994. The scandal of poor medical research. *BMJ* (*British Medical Journal*) 1994; 308:283, doi: https://doi.org/10.1136/bmj.308.6924.283.

Altman, D.G., and Bland, J.M. 1983. Measurement in medicine: the analysis of method comparison studies. *The Statistician* 32: 307-317.

Altman, D.G., and Bland, J.M. 1987. Comparing methods of measurement. *Applied Statistics* 36: 224-225.

Atkinson G., and Nevill, A.M. 1998. Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. *Sports Medicine* 26: 217-238.

Armitage P, Berry, G., and Matthews, J.N.S. 2002. *Statistical methods in medical research*. 4th ed., Blackwell Science, Blackwell Publishing, Malden, Massachusetts, USA.

Aronoff, S. 1982. Classification accuracy: a user approach. *Photogrammetric Engineering and Remote Sensing* 48: 1299-1307.

Aronoff, S. 1985. The minimum accuracy value as an index of classification accuracy. *Photogrammetric Engineering and Remote Sensing* 51: 99-111.

Babu, G.J., and Feigelson, E.D. 1992. Analytical and Monte Carlo comparisons of six different linear least squares fits. *Communications in Statistics - Simulation and Computation* 21(2): 533-549.

Barnhart, H.X. 2018. A review on assessing agreement. *Wiley StatsRef: Statistics Reference Online*, 1-30.

Bater, C.W., Wulder, M.A., Coops, N.C., Nelson, R.F., Hilker, T., and Næsset, E. 2011. Stability of sample-based scanning-LiDAR-derived vegetation metrics for forest monitoring. *Transactions on Geoscience and Remote Sensing* 49: 2385–2392.

Bland, J.M., and Altman, D.G. 1986. Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet* 327**:** 307–310.

Broemeling, L.D. 2009. *Bayesian methods for measures of agreement*. Chapman & Hall/CRC Press, New York. 340 p.

Bunce, C. 2009. Correlation, agreement, and Bland–Altman analysis: statistical analysis of method comparison studies. *American Journal of Ophthalmology* 148: 4-6.

Carstensen, B. 2010. *Comparing clinical measurement methods: a practical guide*. John Wiley & Sons, New York. 172 p.

Chalmers, I., and Glasziou, P. 2009. Avoidable waste in the production and reporting of research evidence. *Lancet* 374: 86-89.

Chicco D., and Jurman G. 2020. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* 21 (1): 6-1–6-13. doi:10.1186/s12864-019-6413-7. PMC 6941312. PMID 31898477.

Chicco D., Toetsch N., and Jurman G. 2021. The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. *BioData Mining* 14 (13): 1-22. doi:10.1186/s13040-021-00244-z. PMC 7863449. PMID 33541410.

Choudhary, P.K., and Nagaraja, H.N. 2005. Measuring agreement in method comparison studies—a review. In *Advances in Ranking and Selection, Multiple Comparisons, and Reliability: Methodology and Applications*, N. Balakrishnan, N. Kannan, and H.N. Nagaraja (editors), pp. 215–244. Birkhäuser, Boston.

Choudhary, P.K., and Nagaraja, H.N. 2017. *Measuring agreement: models, methods, and applications*. John Wiley and Sons, Hoboken, NJ. 336 p.

Cochran, W.G. 1954. Some methods for strengthening the common chi squared tests. *Biometrics* 10: 417-451.

Cohen, J.A. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20(1): 37–46.

Cohen, J.A. 1968. Weighed kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin* 70 (4): 213–220.

Congalton, R.G. 1991. A review of assessing the accuracy of classifications of remotely sensed data. *Remote Sensing of Environment* 37: 35–46.

Congalton, R.G., Oderwald, R.G., and Mead, R.A. 1983. Assessing Landsat classification accuracy using discrete multivariate statistical techniques. *Photogrammetric Engineering and Remote Sensing* 49(12): 1671-1678.

Coops, N.C., Hilker, T., Wulder, M.A., St-Onge, B., Newnham, G., Siggins, A., and Trofymow, J.A. 2007. Estimating canopy structure of Douglas-fir forest stands from discrete-return LiDAR. *Trees* 21: 295-310.

Coops, N.C, Achim, A., Arp, P., Bater, C.W., Caspersen, J.P., Côté, J., Dech, J.P., Dick, A.R., van Ewijk, K., Fournier, R., Goodbody, T.R.H., Hennigar, C.R., Leboeuf, A., van Lier, O.R., Luther, J.E., MacLean, D.A., McCartney, G., Pelletier, G., Prieur, J., Tompalski, P., Treitz, P.M., White, J.C., and Woods, M.E. 2021. Advancing the application of remote sensing

for forest information needs in Canada: lessons learned from a national collaboration of university, industrial and government stakeholders. *The Forestry Chronicle* 97(2): 109-126.
Cortini, F., Comeau, P.G., Strimbu, V.C., Hogg, E.H, Bokalo M., and Huang, S. 2017. Survival functions for boreal tree species in northwestern North America. *Forest Ecology and Management* 402: 177-185.
Davenport, E., and El-Sanhury, N. 1991. Phi/Phimax: review and synthesis. *Educational and Psychological Measurement* 51: 821–828.
Dransfield, R.D., and Brightwell, R. 2012. Avoiding and detecting statistical malpractice: design & analysis for biologists, with R. Published by *InfluentialPoints*, United Kingdom. Online access (cost £25, https://influentialpoints.com/course/).
Draper, N.R., and Smith, H. 1998. *Applied regression analysis*. 3rd edition, John Wiley and Sons, New York. 706 p.
Dunn, G. 2004. *Statistical evaluation of measurement errors: design and analysis of reliability studies*. 2nd ed., Oxford University Press Inc., New York (and Arnold, London). 224 p.
Duveiller, G., Fasbender, D., and Meroni, M. 2016. Revisiting the concept of a symmetric index of agreement for continuous datasets. *Scientific Report* 6, 19401 (*Nature.com*); doi: 10.1038/srep19401. 14 p.
Engsted, T. 2009. Statistical vs. economic significance in economics and econometrics: further comments on McCloskey and Ziliak. *Journal of Economic Methodology* 16(4): 393-408.
Erasmi, S., Semmler, M., Schall, P., and Schlund, M. 2019. Sensitivity of bistatic TanDEM-X data to stand structural parameters in temperate forests. *Remote sensing*, *11,* 2966, 1-18, doi: 10.3390/rs11242966.
Fashi, A., Tsegaye, T., Tadesse, W., and Coleman, T. 2000. Incorporation of digital elevation models with Landsat-TM data to improve land cover classification accuracy. *Forest Ecology and Management* 128: 57−64.
Fawcett, T. 2006. An introduction to ROC analysis. *Pattern Recognition Letters* 27(8): 861–874.
Feinstein, A.R., and Cicchetti, D.V. 1990. High agreement but low Kappa: I. the problems of two paradoxes. *Journal of Clinical Epidemiology* 43(6): 543-549.
Feyerabend, P. 1981. How to defend society from science. In: *Scientific Revolutions*, Ian Hacking (ed.), Oxford readings in philosophy, Oxford University Press, Oxford, U.K.
Finn, J.T. 1993. Use of the average mutual information index in evaluating classification error and consistency. *International Journal of Geographical Information Systems* 7(4): 349-366, DOI: 10.1080/02693799308901966 (published online: 01 Feb 2007).
Fitzgerald, R.W., and Lees, B.W. 1994. Assessing the classification accuracy of multiresource remote sensing data. *Remote Sensing of Environment* 47: 362-368.
Fleiss, J.L. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, Vol. 76, No. 5, pp. 378–382.
Fleiss, J.L., Cohen, J., and Everitt, B.S. 1969. "Large sample standard errors of kappa and weighted kappa". *Psychological Bulletin* 72 (5): 323–327.
Fleiss, J.L., Levin, B., and Paik, M.C. 2003. *Statistical methods for rates and proportions*. 3rd edition, Hoboken, New Jersey: Wiley. https://doi.org/10.1002/0471445428.
Foody, G.M. 1992. On the compensation for chance agreement in image classification accuracy assessment. *Photogrammetric Engineering and Remote Sensing* 58: 1459-1460.
Foody, G.M., 2002. Status of land cover classification accuracy assessment. *Remote Sensing of Environment* 80: 185–201. https://doi.org/10.1016/S0034-4257(01)00295-4.
Foody, G.M. 2004. Thematic map comparison: evaluating the statistical significance of differences in classification accuracy. *Photogrammetric Engineering & Remote Sensing* 70: 627-633.
Foody, G.M. 2020. Explaining the unsuitability of the kappa coefficient in the assessment and comparison of the accuracy of thematic maps obtained by image classification. *Remote Sensing of Environment* 239, 111630. https://doi.org/10.1016/j.rse.2019.111630
Forsite Consultants Ltd. 2020. LiDAR enhanced forest inventory – pilot project. Final report (version 2, April 24, 2020). Prepared for Canadian Forest Products Ltd., Grande Prairie, Alberta, Canada.
Fowlkes, E.B., and Mallows, C.L. 1983. A method for comparing two hierarchical clusterings. *Journal of the American Statistical Association* 78 (383): 553–569.
Freedman, D.H. 2010. Lies, damned lies, and medical science. *The Atlantic*, November 2010 (https://www.theatlantic.com/magazine/archive/2010/11/lies-damned-lies-and-medical-science/308269/).
Freeman, G.H., and Halton, J.H. 1951. Note on an exact treatment of contingency, goodness of fit, and other problems of significance. *Biometrika* 38: 141–149.
Fung, T., and E. LeDrew. 1988. The determination of optimal threshold levels for change detection using various accuracy indices. *Photogrammetric Engineering and Remote Sensing* 54(10): 1449-1454.
Gergel, S.E., Stange, Y., Coops, N.C., Johansen, K., and Kirby, K.R. 2007. What is the value of a good map? an example using high spatial resolution imagery to aid riparian restoration. *Ecosystems* 10: 688–702. https://doi.org/10.1007/s10021-007-9040-0.

Ginevan, M.E. 1979. Testing land use map accuracy: another look. *Photogrammetric Engineering and Remote Sensing* 45(10): 1371-1377.

Glasziou, P., and Chalmers, I. 2016. Is 85% of health research really "wasted"? *The BMJ Opinion*. https://blogs.bmj.com/bmj/2016/01/14/paul-glasziou-and-iain-chalmers-is-85-of-health-research-really-wasted/.

Gwet, K.L. 2012. Handbook of inter-rater reliability: the definitive guide to measuring the extent of agreement among multiple raters. 3rd ed., Advanced Analytics, LLC, Gaithersburg, Maryland. 294 p.

Harrison, S.R. 1990. Regression of a model on real-system output: an invalid test of model validity. *Agricultural Systems* 34: 183–190.

Helldén, U.A. 1980. A test of landsat-2 imagery and digital data for thematic mapping illustrated by an environmental study in northern Kenya, Natural Geography Institute Report, Vol. 47, Lund University, Sweden.

Heurich, M., Persson, Å., Holmgren, J., and Kennel, E. 2004. Detecting and measuring individual trees with laser scanning in mixed mountain forest of central Europe using an algorithm developed for Swedish boreal forest conditions. *International Archives of Photogrammetry*, *Remote Sensing and Spatial Information Sciences*, Vol. XXXVI - 8/W2: 307-312.

Hollis, S. 1996. Analysis of method comparison studies [Guest Editorial]. *Annals of Clinical Biochemistry* 33: 1–4.

Hologa, R., Scheffczyk, K. , Dreiser, C., and Gärtner, S. 2021. Tree species classification in a temperate mixed mountain forest landscape using random forest and multiple datasets. *Remote Sensing* 13, 4657. https://doi.org/10.3390/rs13224657.

Huang, S. 1994. Individual tree volume estimation procedures for Alberta: methods of formulation and statistical foundations. Land and Forest Service, Alberta Environmental Protection Tech. Rep. Pub. No. T/288, Edmonton, Alberta. 80 p.

Huang, S. 2002. Validating and localizing growth and yield models: procedures, problems and prospects. An invited keynote presentation given at "Reality, Models and Parameter Estimation", Sesimbra, Portugal, June 2-5, 2002, jointly sponsored by IUFRO (4.01 and 4.11), Instituto Superior de Gestão and Instituto Superior de Agronomia, Portugal.

Huang, S. 2016. Individual tree diameter prediction models from tree height. Forest Management Branch, Alberta Agriculture and Forestry, Edmonton, Alberta, Technical Report Pub. No. T/606. 59 p.

Huang, S., Meng, S.X., and Yang, Y. 2009a. Prediction implications of nonlinear mixed-effects forest biometric models estimated with a generalized error structure. *In* Proceedings of Joint Statistical Meetings, **Section on Statistics and the Environment**, American Statistical Association, August 1-6, Washington, D.C. pp. 1174–1188.

Huang, S., Meng, S.X., and Yang, Y. 2009b. Assessing the goodness-of-fit of forest models estimated by nonlinear mixed model methods. *Canadian Journal of Forest Research* 39: 2418–2436.

Huang, S., and Titus, S.J. 1999. Estimating a system of nonlinear simultaneous individual tree models for white spruce in boreal mixed-species stands. *Canadian Journal of Forest Research* 29: 1805-1811.

Huang, S., Titus, S.J., Price, D. and Morgan, D.J. 1999. Validation of ecoregion-based taper equations for white spruce in Alberta. *The Forestry Chronicle* 75(2): 281-292.

Huang, S., Wiens, D.P., Yang, Y., Meng, S.X., and VanderSchaaf C.L. 2009c. Assessing the impacts of species composition, top height and density on individual tree height prediction of quaking aspen in boreal mixedwoods. *Forest Ecology and Management* 258: 1235-1247.

Huang, S., Yang, Y., and Aitkin, D. 2013. Population and plot-specific individual tree height-diameter models for major Alberta tree species. Alberta Environment and Sustainable Resource Development, Edmonton, Alberta, Technical Report Pub. No. T/600. 81 p.

Huang, S., Yang, Y., and Aitkin, D. 2016. Updated population and plot-specific individual tree height-diameter models for major Alberta tree species. Forest Management Branch, Alberta Agriculture and Forestry, Edmonton, Alberta, Technical Report Pub. No. T/609. 76 p.

Huang, S., Yang, Y. and Meng, S.X. 2011. Developing forest models from longitudinal data: a case study assessing the need to account for correlated and/or heterogeneous error structures under a nonlinear mixed model framework. *Journal of Forest Planning* 16: 121–131.

Huang, S., Zaichkowsky, M., Weeks D., Li, C., Brown, C., Parlow, M., Buckmaster, G., Tansanu, C., and Yang, Y. 2019. Method comparison and method calibration applicable to forest measurements and model predictions. Forest Stewardship and Trade Branch, Alberta Agriculture and Forestry, Edmonton, Alberta, Technical Report Pub. No.: T/2019–RA01. 126 p. https://open.alberta.ca/publications/9781460143759.

Hudson, W., and Ramm, C. 1987. Correct formulation of the kappa coefficient of agreement. *Photogrammetric Engineering and Remote Sensing* 53(4):421-422.

Hurlbert, S.H., and Lombardi, C.M. 2003. Design and analysis: uncertain intent, uncertain result. Book review of Quinn, G.P. & Keough, M.J. (2002). Experimental design and data analysis for biologists. CUP, New York. *Ecology* 84(3): 810-812.

Ioannidis, J.P.A. 2005. Why most published research findings are false. *Public Library of Science* (*PLOS*) *Medicine* 2(8): e124. doi:10.1371/journal.pmed.0020124.

Isobe, T., Feigelson, E.D., Akritas, M.G., and Babu, G.J. 1990. Linear regression in astronomy I. *Astrophysical Journal* 364: 104-113.

Ji, L., and Gallo, K. 2006. An agreement coefficient for image comparison. *Photogrammetric Engineering and Remote Sensing* 72: 823–833.

Ji, L., Gallo, K., Eidenshink, J.C., and Dwyer, J. 2008. Agreement evaluation of AVHRR and MODIS 16-day composite NDVI data sets. *International Journal of Remote Sensing* 29: 4839-4861.

Ke, Y., and Quackenbush, L.J. 2011. A review of methods for automatic individual tree-crown detection and delineation from passive remote sensing. *International Journal of Remote Sensing* 32: 4725–4747.

Kleijnen, J.P.C. 1999. Validation of models: statistical techniques and availability. *In* Proceedings of the 1999 Winter Simulation Conference, 5–8 December 1999, Phoenix, Ariz. *Edited by* P.P. Farrington, H.B. Nembhard, D.T. Sturrock, and G.W. Evans. Institute of Electrical and Electronics Engineers, New York. pp. 647–654.

Kleijnen, J.P.C., Bettonvil, B., and Groenendaal, W.V. 1998. Validation of trace-driven simulation models: a novel regression test. *Management Science* 44: 812–819.

Koukoulas, S., and Blackburn, G.A. 2001. Introducing new indices for accuracy evaluation of classified images representing semi-natural woodland environments. *Photogrammetric Engineering and Remote Sensing* 67(4): 499-510.

Krause, S., Sanders, T.G.M., Mund, J.P., and Greve K. 2019. UAV-based photogrammetric tree height measurement for intensive forest monitoring. *Remote Sensing* 11, 758; doi:10.3390/rs11070758.

Krippendorff, K. 1978. Reliability of binary attribute data. *Biometrics* 34(1): 142–144.

Krippendorff, K. 2018. *Content analysis: an introduction to its methodology*. 4th ed., Sage Publications, Inc., Thousand Oaks, CA. 472 pages.

Krummenauer, F., and Doll, G. 2000. *Statistical* methods for the *comparison* of measurements derived from orthodontic imaging. *European Journal of Orthodontics 22: 257-269.*

Kvålseth, T.O. 2017. On normalized mutual information: measure derivations and properties. *Entropy* 2017, 19, 631; doi:10.3390/e19110631.

Kwak, D.A., Lee, W.K., Lee, J.H., Biging, G.S., and Gong, P. 2007. Detection of individual trees and estimation of tree height using LiDAR data. *Journal of Forest Research* 12: 425–434.

Lambdin, C. 2012. Significance tests as sorcery: science is empirical—significance tests are not. *Theory & Psychology* 22: 67-90.

Legates, D.R., and McCabe, G.J. Jr. 1999. Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation. *Water Resources Research* 35**:** 233–241.

Li, Y., Andersen, H.E., and McGaughey, R. 2008. A comparison of statistical methods for estimating forest biomass from light detection and ranging data. *Western Journal of Applied Forestry* 23: 223–231.

Liao, J.J.Z. 2003. An improved concordance correlation coefficient. *Pharmaceutical Statistics* 2: 253-261.

Liao, J.J.Z., and Capen, R.C. 2009. Multiple evaluators. In *Wiley Encyclopedia of Clinical Trials*, Eds. R.B. D'Agostino, L.M. Sullivan, and J.M. Massaro, vol. 3, pp. 186–194, Wiley, Hoboken, New Jersey, USA.

Liao, J.J.Z., and Lewis, J.W. 2000. A note on concordance correlation coefficient. *PDA Journal of Pharmaceutical Science & Technology* 54: 23-26.

Lillesand, T., Kiefer, R.W., and Chipman, J. 2015. *Remote sensing and image interpretation*. 7th edition, John Wiley & Sons, Hoboken, New Jersey, USA.

Lim, K., Treitz, P., Groot, A., and St-Onge, B. 2001. Estimation of individual tree heights using LIDAR remote sensing. *Proceedings of the twenty-third annual Canadian symposium on remote sensing*, Quebec, QC, August 20-24, 2001 (CD-ROM).

Lim, K., Treitz, P., Wulder, M.A., St-Onge, B., and Flood, M. 2003. LiDAR remote sensing of forest structure. *Progress in Physical Geography* 27: 88–106.

Lin, L.I. 1989. A concordance correlation coefficient to evaluate reproducibility. *Biometrics* 45: 255-268.

Lin, L., Hedayat, A.S., Sinha, B., and Yang, M. 2002. Statistical methods in assessing agreement: models, issues, and tools. *Journal of the American Statistical Association* 97: 257-270.

Lin, L.I., Hedayat, A.S., and Wu, W. 2012. *Statistical tools for measuring agreement*. Springer, New York. 161 p.

Liu, C., Fraizer, P., and Kumar, L. 2007. Comparative assessment of the measures of thematic classification accuracy. *Remote Sensing of Environment* 107: 606-616.

Liu, G., Wang, J., Dong, P., Chen, Y., and Liu, Z. 2018. Estimating individual tree height and diameter at breast height (DBH) from terrestrial laser scanning (TLS) data at plot level. *Forests* (2018) 9: 398. doi:10.3390/f9070398.

Lose, G., and Klarskov, N. 2017. Why published research is untrustworthy. *International Urogynecology Journal* 28: 1271–1274.

Ludbrook, J. 2002. Statistical techniques for comparing measurers and methods of measurement: a critical review. *Clinical and Experimental Pharmacology and Physiology* 29: 527-536.

Ma, Z., and Redmond, R. L. 1995. Tau coefficients for accuracy assessment of classification of remote sensing data. *Photogrammetric Engineering and Remote Sensing* 61: 435-439.

Makin, T.R., and Orban de Xivry, J.-J. 2019. Ten common statistical mistakes to watch out for when writing or reviewing a manuscript. *eLife* 8, e48175.

Matasci, G., Hermosilla, T., Wulder, M.A., White, J.C., Coops, N.C., Hobart, G.W., and Zald, H.S.J. 2018. Large-area mapping of Canadian boreal forest cover, height, biomass and other structural attributes using Landsat composites and lidar plots. *Remote Sensing of Environment* 209: 90-106.

Matthews, B.W. 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure* 405 (2): 442-451.

Means, J.E., Acker, S.A., Harding, D.J., Blair, J.B., Lefsky, M.A., Cohen, W.B., Harmon, M.E., and McKee, W.A. 1999. Use of large-footprint scanning airborne Lidar to estimate forest stand characteristics in the Western Cascade of Oregon. *Remote Sensing of Environment* 67: 298–308.

Mehta, C.R., and Patel, N.R. 1983. A network algorithm for performing Fisher's exact test in contingency tables. *Journal of the American Statistical Association* 78: 427–434.

Meng, S.X., and Huang, S. 2009. Improved calibration of nonlinear mixed-effects models demonstrated on a height growth function. *Forest Science* 55: 238-248.

Meng, S.X. and Huang, S. 2010. Incorporating correlated error structure into mixed forest growth models: prediction and inference implications. *Canadian Journal of Forest Research* 40: 977-990.

Meng, S.X., Huang, S., VanderSchaaf, C.L., Yang, Y. and Trincado, G. 2012. Accounting for serial correlation and its impact on forecasting ability of a fixed- and mixed-effects basal area model: a case study. *European Journal of Forest Research* 131: 541–552.

Mielke, P.W., Jr. 1984. Meteorological applications of permutation techniques based on distance functions, *Handbook of Statistics,* Vol. 4, pp. 813–830, Elsevier, Amsterdam, The Netherlands.

Mielke, P.W., Jr. 1991. The application of multivariate permutation methods based on distance functions in the earth sciences. *Earth-Science Reviews* 31: 55–71.

Mohan, M., Silva, C.A., Klauberg, C., Jat, P., Catts, G., Cardil, A., Hudak, A.T., and Dia, M. 2017. Individual tree detection from unmanned aerial vehicle (UAV) derived canopy height model in an open canopy mixed conifer forest. *Forests* 2017, 8, 340; doi:10.3390/f8090340.

Monserud, R.A., Huang, S., and Yang, Y. 2006. Predicting lodgepole pine site index from climatic parameters in Alberta. *The Forestry Chronicle* 82(4): 562–571.

Müller, R., and Büttner, P. 1994. A critical discussion of intraclass correlation coefficients. *Statistics in Medicine* 13: 2465–2476.

Nakai, T., Sumida, A., Kodama, Y., Hara, T., and Ohta, T. 2010. A comparison between various definitions of forest stand height and aerodynamic canopy height. *Agricultural and Forest Meteorology* 150(9): 1225-1233.

Neeti, N., and Kennedy, R. 2016. Comparison of national level biomass maps for conterminous US: understanding pattern and causes of differences. *Carbon Balance and Management* 11:19, DOI 10.1186/s13021-016-0060-y.

Nelson, R.F. 1983. Detecting forest canopy change due to insect activity using Landsat MSS. *Photogrammetric Engineering and Remote Sensing* 49(9): 1303-1314.

Neter, J., Wasserman, W., and Kutner, M.H. 1989. *Applied linear regression models* (2nd ed.), Richard D. Irwin, Inc., Homewood, IL. 688 p.

Nuzzo R. 2014. Statistical errors. *Nature* 506: 150–152.

Næsset, E. 1997. Determination of mean tree height of forest stands using airborne laser scanner data. *ISPRS Journal of Photogrammetry and Remote Sensing* 52: 49–56.

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: 88–99.

Næsset, E., Bollandsås, O.M., and Gobakken, T. 2005. Comparing regression methods in estimation of biophysical properties of forest stands from two different inventories using laser scanner data. *Remote Sensing of Environment* 94: 541–553.

Næsset, E., Gobakken, T., Bollandsås, O.M., Gregoire, T.G., Nelson, R., and 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.

Olofsson, P., Foody, G.M., Herold, M., Stehman, S.V., Woodcock, C.E., and Wulder, M.A. 2014. Good practices for estimating area and assessing accuracy of land change. *Remote Sensing of Environment* 148: 42-57.

Piñeiro, G., Perelman, S., Guerschman, J.P., and Paruelo, J.M. 2008. How to evaluate models: observed vs. predicted or predicted vs. observed? *Ecological Modelling* 216: 316–322.

Pontius, R.G., Jr., and Santacruz, A. 2014. Quality, exchange, and shift components of difference in a square contingency table. *International Journal of Remote Sensing* 35: 7543-7554.

Powers, D.M.W. 2011. Evaluation: from precision, recall and f-measure to roc, informedness, markedness & correlation. *Journal of Machine Learning Technologies* 2(1): 37–63.

Powers, D.M.W. 2012. The problem with kappa. Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, Avignon, France, April 23 - 27, pp: 345–355.

Prieur, J.F., St-Onge, B., Fournier, R.A., Woods, M.E., Rana, P., and Kneeshaw, D. 2022. A comparison of three airborne laser scanner types for species identification of individual trees. *Sensors* 22(1), 35; https://doi.org/10.3390/s22010035.

Puliti, S., Gobakken, T., Ørka, H.O., and Næsset, E. 2017. Assessing 3D point clouds from aerial photographs for species-specific forest inventories. *Scandinavian Journal of Forest Research* 32: 68–79.

Radoux, J., and Bogaert, P. 2017. Good practices for object-based accuracy assessment. *Remote Sensing* 9, 646. https://doi.org/10.3390/rs9070646.

Riemann, R., Wilson, B.T., Lister, A., and Parks, S. 2010. An effective assessment protocol for continuous geospatial datasets of forest characteristics using USFS Forest Inventory and Analysis (FIA) data. *Remote Sensing of Environment* 114: 2337–2352.

Robinson, W.S. 1957. The statistical measurement of agreement. *American Sociological Review* 22: 17–25.

Rosenfield, G.H., and Fitzpatrick-Lins, K.A. 1986. A coefficient of agreement as a measure of thematic classification accuracy. *Photogrammetric Engineering and Remote Sensing* 52(2): 223-227.

Sammut, C., and Webb, G.I. (eds.). 2011. Encyclopedia of machine learning (2010th edition, March 28 2011 publishing), Springer, New York. 1031 pp. doi:10.1007/978-0-387-30164-8.

SAS Institute Inc. 2011. SAS/QC 9.3 User's Guide, SAS Institute Inc., Cary, NC. 2323 p.

SAS Institute Inc. 2020. SAS Viya programming documentation 2020.1.4, SAS Institute Inc., Cary, NC, USA.

Schaefer, J.T. 1990. The critical success index as an indicator of warning skill. *Weather and Forecasting* 5: 570–575.

Schober, P., Bossers, S.M., and Schwarte, L.A. 2018. Statistical significance versus clinical importance of observed effect sizes: what do *p* values and confidence intervals really represent? *Anesthesia & Analgesia* 126(3): 1068-1072.

Scofield, G.B., Pantaleão, E., and Negri, R.G. 2015. A comparison of accuracy measures for remote sensing image classification: case study in an Amazonian region using support vector machine. *International Journal of Image Processing* 9(1): 11-21.

Scott, W.A. 1955. Reliability of content analysis: the case of nominal scale coding. *Public Opinion Quarterly* 19(3): 321-325.

Short, N.M. 1982. The landsat tutorial workbook - basics of satellite remote sensing, Greenbelt, Md, Goddard Space Flight Center, NASA reference publication 1078.

Shoukri, M.M. 2010. *Measures of interobserver agreement and reliability*. 2nd ed., Chapman & Hall/CRC Press, New York. 291 p.

Sibona, E., Vitali, A., Meloni, F., Caffo, L., Dotta, A., Lingua, E., Motta, R., and Garbarino, M. 2017. Direct measurement of tree height provides different results on the assessment of LiDAR accuracy. *Forests* 8, 7; doi:10.3390/f8010007.

Smith, R. 2014. Medical research - still a scandal. *The BMJ Opinion*, https://blogs.bmj.com/bmj/2014/01/31/richard-smith-medical-research-still-a-scandal/.

Smits, P.C., Dellepiane, S.G., and Schowengerdt, R.A. 1999. Quality assessment of image classification algorithms for land-cover mapping: A review and a proposal for a cost-based approach. *International Journal of Remote Sensing* 20: 1461−1486.

Stehman, S.V. 1997. Selecting and interpreting measures of thematic classification accuracy. *Remote Sensing of Environment* 62(1): 77–89.

Stehman, S.V. 2004. A critical evaluation of the normalized error matrix in map accuracy assessment. *Photogrammetric Engineering and Remote Sensing* 70: 743−756.

Stehman, S.V., and Czaplewski, R.L. 1998. Design and analysis for thematic map accuracy assessment: fundamental principles. *Remote Sensing of Environment* 64: 331−344.

Stehman, S.V., and Foody, G.M. 2019. Key issues in rigorous accuracy assessment of land cover products. *Remote Sensing of Environment*, Vol 231 (111199), https://doi.org/10.1016/j.rse.2019.05.018.

Stokes, M.E., Davis, C.S., and Koch, G.G. 2012. *Categorical data analysis using SAS*. 3rd ed. SAS Institute Inc., Cary, NC, USA.

St-Onge, B., Treitz, P., and Wulder, M.A. 2003. Tree and canopy height estimation with scanning LiDAR. In: *Remote sensing of forest environments: concepts and case studies* (Wulder, M.A. and Franklin, S.E., eds.), pp.489–509. Kluwar, Boston.

Story, M., and R.G. Congalton. 1986. Accuracy assessment: a user's perspective. *Photogrammetric Engineering and Remote Sensing* 52(3): 397-399.

Strehl, A., and Ghosh, J. 2002. Cluster ensembles - a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research* 3: 583-617.

Strîmbu, V.F., and Strîmbu, B.M. 2015. A graph-based segmentation algorithm for tree crown extraction using airborne LiDAR data. *ISPRS Journal of Photogrammetry and Remote Sensing* 104: 30–43.

Surový, P., and Kuželka, K. 2019. Acquisition of forest attributes for decision support at the forest enterprise level using remote-sensing techniques—a review. *Forests* 10(3), 273.

Tharwat, A. 2020. Classification assessment methods. *Applied Computing and Informatics*. Volume 17, Issue 1. doi:10.1016/j.aci.2018.08.003, July 30, 2020.

Türk, G. 1979. GT index: A measure of the success of predictions. *Remote Sensing of Environment* 8: 65−75.

Türk, G. 2002. Letter to the editor: map evaluation and “chance correction”. *Photogrammetric Engineering and Remote Sensing* 68, 123−129.

Ungerer, J.P.J., and Pretorius, C.J. 2017. Method comparison - a practical approach based on error identification. *Clinical Chemistry and Laboratory Medicine* 56 (1): 1-4.

Ustin, S.L., Hart, Q.J., Duan, L., and Scheer, G. 1996. Vegetation mapping on hardwood rangelands in California. *International Journal of Remote Sensing* 17: 3015−3036.

Van Huffel, S. (Ed.). 1997. Recent advances in total least squares techniques and errors-in-variables modeling. *SIAM Proceedings Series*, SIAM, Philadelphia. 377p.

Van Noorden, R., Maher, B., and Nuzzo, R. 2014. The top 100 papers. *Nature* 514: 550-553.

Vastaranta, M., Kankare, V., Holopainen, M., Yu, X., Hyyppä, J., and Hyyppä, H. 2012. Combination of individual tree detection and area-based approach in imputation of forest variables using airborne laser data. *ISPRS Journal of Photogrammetry and Remote Sensing* 67:73–79.

von Eye, A., and Mun, E.Y. 2004. *Analyzing rater agreement: manifest variable methods*. Psychology Press, New York. 202 p.

Wang, Y., Lehtomaki, M., Liang, X., Pyorala, J., Kukko, A., Jaakkola, A., Liu, J., Feng, Z., Chen, R., and Hyyppa, J. 2019. Is field-measured tree height as reliable as believed – a comparison study of tree height estimates from field measurement, airborne laser scanning and terrestrial laser scanning in a boreal forest. *ISPRS Journal of Photogrammetry and Remote Sensing* 147: 132-145.

Wasserstein, R.L., and Lazar, N.A. 2016. The ASA's statement on p-values: context, process, and purpose. *The American Statistician* 70(2): 129–133.

Watterson, I.G. 1996. Non-dimensional measures of climate model performance. *International Journal of Climatology* 16: 379–391.

White, J.C., Tompalski, P., Vastaranta, M., Wulder, M.A., Saarinen, N., Stepper, C., and Coops, N.C. 2017. A model development and application guide for generating an enhanced forest inventory using airborne laser scanning data and an area-based approach. Natural Resources Canada, Canadian Forest Service, Information Rep. FI-X-018, Victoria, British Columbia. 38 p.

Willmott, C.J. 1981. On the validation of models. *Physical Geography* 2: 184–194.

Willmott, C.J. 1982. Some comments on the evaluation of model performance. *Bulletin American* Meteorological Society 63: 1309–1313.

Willmott, C.J., Ackleson, S.G., Davis, R.E., Feddema, J.J., Klink, K.M., Legates, D.R., O’Donnell, J., and Rowe, C.M. 1985. Statistics for the evaluation and comparison of models. *Journal of Geophysical Research* 90 (C5): 8995-9005.

Wilson, B.T., Lister, A.J., and 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.

Wu, B., Yu, B., Wu, Q., Huang, Y., Chen, Z., and Wu, J. 2016. Individual tree crown delineation using localized contour tree method and airborne LiDAR data in coniferous forests. *International Journal of Applied Earth Observation and Geoinformation* 52: 82–94.

Yang, W, and Kondoh, A. 2020. Evaluation of the Simard et al. 2011 global canopy height map in boreal forests. *Remote Sensing* 12(7), 1114, https://doi.org/10.3390/rs12071114.

Yang, Y., Monserud, R.A., and Huang, S. 2004. An evaluation of diagnostic tests and their roles in validating forest biometric models. *Canadian Journal of Forest Research* 34: 619–629.

Yang, Y., and Huang, S. 2011a. Estimating a multilevel dominant height-age model from nested data with generalized errors. *Forest Science* 57:102–116.

Yang, Y., and Huang, S. 2011b. Comparison of different methods for fitting nonlinear mixed forest models and for making predictions. *Canadian Journal of Forest Research* 41: 1671-1686.

Yang, Y., and Huang, S. 2013. A generalized mixed logistic model for predicting individual tree survival probability with unequal measurement intervals. *Forest Science* 59(2): 177–187.

Yang, Y., and Huang, S. 2014. Suitability of five cross validation methods for performance evaluation of nonlinear mixed-effects forest models – a case study. *Forestry* 87: 654–662.

Yang, Y., and Huang, S. 2015. Two-stage ingrowth models for four major tree species in Alberta. *European Journal of Forest Research* 134: 991–1004.

Yang, Y., Titus, S.J. and Huang, S. 2003. Modeling individual tree mortality for white spruce in Alberta. *Ecological Modelling* 163(3): 209–222.

Yu, X., Hyyppä, J., Vastaranta, M., Holopainen, M., and Viitala, R. 2011. Predicting individual tree attributes from airborne laser point clouds based on the random forests technique. *ISPRS Journal of Photogrammetry and Remote Sensing* 66: 28–37.

Zald, H.S.J., Wulder, M.A., White, J.C., Hilker, T., Hermosilla, T., Hobart, G.W., and Coops, N.C. 2016. Integrating Landsat pixel composites and change metrics with lidar plots to predictively map forest structure and aboveground biomass in Saskatchewan, Canada. *Remote Sensing of Environment* 176: 188-201.

Zhang, J., Hu, J., Lian, J., Fan, Z., Ouyang, X., Ye, W. 2016. Seeing the forest from drones: testing the potential of lightweight drones as a tool for long-term forest monitoring. *Biological Conservation* 198: 60–69.

Zhang, J., Zhang, Z., Lutz, J.A., Chu, C., Hu, J., Shen, G., Li, B., Yang, Q., Lian, J., Zhang, M., Wang, X., Ye, W., and He, F. 2022. Drone-acquired data reveal the importance of forest canopy structure in predicting tree diversity. *Forest Ecology and Management* 505 (6): 119945. https://doi.org/10.1016/j.foreco.2021.119945.

Zhao, K., Suarez, J.C., Garcia, M., Hu, T., Wang, C., and Londo, A. 2018. Utility of multitemporal lidar for forest and carbon monitoring: tree growth, biomass dynamics, and carbon flux. *Remote Sensing of Environment* 204: 883-897.

Zhuang, X., Engel, B.A., Xiong, X., and Johannsen, C.J. 1995. Analysis of classification results of remotely sensed data and evaluation of classification algorithms. *Photogrammetric Engineering and Remote Sensing* 61, 427−433.