ARBOR: A New Framework for Assessing the Accuracy of Individual Tree Crown Delineation from Remotely-sensed Data

Jon Murray[1*], David Gullick[1], George Alan Blackburn[1], James Duncan Whyatt[1], and Christopher Edwards[2]

[1]Lancaster Environment Centre, Lancaster University, Lancashire, LA1 4YQ

[2]School of Computing and Communications, Lancaster University, Lancaster, LA1 4WA.

*Corresponding author: Tel: +44 1524 652 01

Email: j.murray3@lancaster.ac.uk

Abstract

To assess the accuracy of individual tree crown (ITC) delineation techniques the same tree needs to be identified in two different datasets, for example, ground reference (GR) data and crowns delineated from LiDAR. Many studies use arbitrary metrics or simple linear-distance thresholds to match trees in different datasets without quantifying the level of agreement. For example, successful match-pairing is often claimed where two data points, representing the same tree in different datasets, are located within 5m of one another. Such simple measures are inadequate for representing the multi-variate nature of ITC delineations and generate misleading measures of delineation accuracy. In this study, we develop a new framework for objectively quantifying the agreement between GR and remotely-sensed tree datasets: the Accuracy of Remotely-sensed Biophysical Observation and Retrieval (ARBOR) framework. Using common biophysical properties of ITC delineated trees (location, height and crown area), trees represented in different data sets were modelled as overlapping Gaussian curves to facilitate a more comprehensive assessment of the level of agreement. Extensive testing quantified the limitations of some frequently used match-pairing methods, in particular, the Hausdorff distance algorithm. We demonstrate that within the ARBOR framework, the Hungarian combinatorial optimisation algorithm improves the match between datasets, while the Jaccard similarity coefficient is effective for measuring the correspondence between the matched data populations. The ARBOR framework was applied to GR and remotely-sensed tree data from a woodland study site to demonstrate how ARBOR can identify the optimum ITC delineation technique, out of four different methods tested, based on two measures of statistical accuracy. Using ARBOR will limit further reliance on arbitrary thresholds as it provides an objective approach for quantifying accuracy in the development and application of ITC delineation algorithms.

Keywords

LiDAR, Individual Tree Crown (ITC), Delineation, Error Detection, Data Matching, Accuracy.

Highlights

1. ARBOR answers the need for a standardised ITC delineation accuracy assessment
2. Similarity of RS-derived and reference trees assessed using biophysical properties
3. Optimised algorithm applied to matching RS-derived and reference tree populations
4. ARBOR quantifies accuracy using biophysical data and data population size
5. ARBOR is a modular framework for the objective assessment of ITC delineations

1.0 Introduction

Individual tree crown (ITC) delineation is an important technique for many environmental remote sensing (RS) studies. These types of investigations include data driven activities such as forest inventories and management, carbon and biomass accounting, tree growth modelling and many other geo-spatial data applications. The ability to accurately delineate individual trees from remotely sensed data is essential for many forest monitoring applications (Eysn, Hollaus et al. 2012, Jakubowksi, Guo et al. 2013, Duncanson, Dubayah et al. 2015, Wu, Yu et al. 2016, Zhen, Quackenbush et al. 2016). ITC delineation, sometimes referred to as tree segmentation, is typically associated with the analysis of high resolution optical imagery or 3D point clouds captured from light detection and ranging (LiDAR). ITC delineation is a process where different methods, often computational and automated, identify high peaks in canopy data as the first step in locating individual trees. This phase is followed by a segmentation procedure, such as watershedding, valley formation or other similar methods, to determine the locations and crown perimeters of individual trees. Typically, to assess the validity of ITC delineation a comparison is made with ground reference (GR) tree data. The comparison requires that individual trees are matched between the two datasets and this pairing is used to assess accuracy of the ITC delineation. In many studies, Euclidean distance is used to pair trees from the different datasets. This has the effect of considering the tree-to-tree matching problem only from a plan perspective, and does not account for tree height or crown area (Yu, Hyyppä et al. 2006, Kwak, Lee et al. 2007, Hladik and Alber 2012, Lu, Guo et al. 2014, Zhen, Quackenbush et al. 2016, Yu, Hyyppä et al. 2017).

Additional insights can be obtained through the combination of ITC delineated trees and other spatial data. For example, canopy height models (CHM) characterise the upper surfaces of the delineated tree crown area and provide opportunities to calculate biophysical properties such as tree height or crown area (Rahman and Gorte 2009). Zhen, Quackenbush et al. (2016) note that validation is a key issue in ITC delineation studies. Typically, validation involves assessment of the outputs of ITC delineation procedures in terms of the precision and accuracy of tree locations and biophysical properties (Leckie, Walsworth et al. 2016). However, there are other issues that complicate the match-pairing ITC delineation, such as the self-optimising growth habits of trees in woodlands (see *supplementary information*). Any resulting ITC delineation anomalies can subsequently lead to the spurious identification of tree crowns (Kwak, Lee et al. 2007), causing the pairing of trees that should not be present in the dataset, or otherwise, through the generation of false-positive matches.

Problems that occur in the match-pairing process are further compounded when analysing data population sizes. A significant consideration when matching pairs of trees is the directionality of the match that is made. Essentially this is the matching of data A to data B in the matching sequence, or, matching data B to data A. Errors that arise from directionality differences can result in the same matches not being achieved in both directions, influenced by the data that is used first as the primary dataset. A solution is bidirectional matching, i.e. matching A-B then B-A, and selecting the best agreement (Singh, Evans et al. 2015). However, this approach reduces the data population as the unmatched trees are unassigned, leading to losses from the dataset. An additional problem is that sorting the order of the data effects match-pairings, as does the order sequence that the algorithm attempts the pairings (Holmgren and Lindberg 2013), for example, matching the tallest trees first. Some data preparation methods sort data by size as part of the processing steps (Kandare, Ørka et al. 2016), however, within tree-to-tree matched-pairing, this may block later trees in the dataset that would have been a more suitable pairing, as the primary tree is already allocated to a corresponding tree. GR data frequently contains many smaller and lower canopy trees that are readily assigned to pairings that are not a suitable match (Holmgren and Lindberg 2013). Trees that are observed in the GR data and not seen in the ITC delineation are data omissions as a product of the data population A, not being the same size as the population B or *vice-versa*. Similarly, commission errors occur where trees are incorrectly assigned to a match-pairing, or assigned to the wrong tree (Holmgren and Lindberg 2013). Typically these errors are related to the ITC delineation method used.

Despite the recognised importance of data validation, in a meta-analysis of 210 studies, only 14.3% validated ITC delineation at a forest stand level, 30% validated ITC delineation on individual trees, and 23.3% at both levels (Zhen, Quackenbush et al. 2016). Significantly, in 32.4% of the studies, no ITC validation was attempted at all. This suggests that there is a pressing need for a standardised method for evaluating the accuracy of ITC delineation techniques, which can be applied widely and consistently (Zhen, Quackenbush et al. 2016). It is also apparent from the literature that no standardised accuracy assessment procedure currently exists, and where ITC delineation techniques have been evaluated this has been on the basis of arbitrary metrics or simple linear distance thresholds. Therefore, there is the need for analytical metrics to quantify the accuracy with which ITC delineations estimate data population size and tree biophysical properties. The research outlined in this paper describes a repeatable and transparent solution for validating ITC delineation techniques that can be applied to individual trees, plots or stands. This paper describes the development of the Assessment of Remotely-sensed Biophysical Observations and Retrieval (ARBOR) framework.

2.0 Aim and Objectives

The aim of this research is to develop a technique for quantifying the accuracy of ITC delineation methods. This requires improving tree-to-tree match-pairing with metrics that include additional analytical parameters beyond simple location or linear distance measurement. Furthermore, metrics are required to find an optimal way in applying the match-pairing to, and achieving the best match for, the overall data population. This approach needs to be robust to the influence of directionality, data order and data omissions. If fulfilled, these requirements allow ITC delineation accuracy in RS data to be assessed in an objective manner. This will be achieved by addressing the following objectives:

1. Identifying a suitable technique for quantifying the similarity of a tree as represented in RS-derived and ground reference datasets, using the biophysical properties: tree location, height and crown area.
2. Determining an optimal algorithm for matching an entire population of trees represented in both RS-derived and ground reference datasets, avoiding introduced bias from directionality, data omissions and other similar factors.
3. Developing metrics for quantifying the accuracy of population size and tree biophysical properties
4. Applying the optimal algorithm and metrics to quantify the accuracy of a variety of ITC delineation methods applied to RS data of a woodland study site.

3.0 Methodology

The methodology for developing the ARBOR framework directly addresses each of the objectives outlined above. Objectives 1-3 will be met by development and testing within a synthetic data environment, to establish the validity of the different analytical elements that will be used within the ARBOR framework. Following the development of the framework and validation of the components that will be used in ARBOR, Objective 4 will be met by applying the ARBOR framework to quantify the match-pairing of real-world data, therefore, providing proof of concept.

3.1 Quantifying the Similarity of a Tree as Represented in RS-derived and Ground Reference Datasets

3.1.1 Defining the Biophysical Properties of a tree.

Jing, Hu et al. (2012) state that differentiation between natural tree crowns is influenced by both the width and depth of the inter-canopy space, in addition to the computationally delineated, circular crown shape. Correspondingly, each tree crown in this study can be considered to have at least a location, height and crown area. It is understood that within broadleaved trees that there may be a linear distance offset between the central point of the stem and the highest green tip of the crown, however, usual forestry conventions are to measure to the highest live point irrespective of any offsetting (West, 2009). To quantify correspondence between two trees, or more specifically, a tree represented in RS-derived data and the same tree in the GR data, the metric criteria has to consider spatial proximity, tree height and overall crown area. Also, for the accuracy comparison to be made on a like-for-like basis, metrics should report successful similarity indices with values of between 0 (impossible) and 1 (certain or identical). Note: In this paper, we have chosen to use GR data as the reference data against which ITC delineations are validated. However, the ARBOR framework can use reference data that has been collected using non-field based methods, such as through manual interpretation of aerial photography.

3.1.2 Limitations of Commonly Used Tree-to-tree Match-pairing Methods

Some tree-to-tree match-pairing agreements are based upon the Euclidean distance between trees (Yu, Hyyppä et al. 2006), however, this approach has problems that may not be adequately resolved. For example, the 2D measurement of the planar distance between the tops of trees assumes that each tree only has a singular apical point. Kaartinen, Hyyppä et al. (2012) note that additional trees in the lower canopy can lead to omission errors between GR and ITC delineated trees. Alternatives consider tree-to-tree pairwise-matching from a 3D model perspective, with linear distance statistics such as the Hausdorff distance algorithm, used to assess the linear correspondence between two points from different datasets (Yu, Hyyppä et al. 2006, Yu, Hyyppä et al. 2017, Zhao, Suarez et al. 2018). The Hausdorff algorithm meets the metric criteria following rescaling the index between 0 and 1, however, due to the distance between the delineated edges of a tree crown, omission errors can occur. Hausdorff can be used in data point comparison, but can be influenced by directionality. To counter this effect, a geometric shape for the crown, such as a circle, has to be used when calculating Hausdorff.

3.2 Gaussian Overlapping and the Jaccard Similarity Coefficient

The analysis of the overlaps between two Gaussian curves (also known as a Gaussian overlap model), measures the comparative distance between the two distributions (Nowakowska, Koronacki et al. 2014). This approach uses the curve centre as the tree location, with the apex indicating the overall tree height and the area under the curve representing the circular crown area. A component overlap analysis of the mixed, normal data distributions identifies changes in the curve location, height and crown area between the overlapping parabolas

(Nowakowska, Koronacki et al. 2015). A Gaussian overlap models where a single tree, identified and described in both datasets, can be aligned to a potential match in the opposing dataset and any similarities in the biophysical properties compared and quantified. Issues regarding complexities in the biophysical properties of trees are discussed further in *supplementary information*.

To satisfy the analysis criteria, the area of overlap between each Gaussian representation of the tree's biophysical properties is assessed. Similar trees achieve greater Gaussian overlap than non-similar trees. To quantify the overlap as a normalised value, the Jaccard similarity coefficient is calculated. Jaccard is the quotient produced by the division of the intersection by the union and measures the observable similarities between two finite data sets. Functionally, Jaccard is a simple measure of the binary distance between data and describes the presence or absence of data, as defined at equation (1).

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \tag{1}$$

A perfect match is a Jaccard value of one, while inferior matches decrease Jaccard towards zero. Due to the infinite nature of the tails on a Gaussian curve, an absolute score of zero cannot be achieved as an inferior score representing a more heavily degenerated match always remains mathematically possible.

Figure 1 uses some examples to demonstrate the Gaussian overlap method and Jaccard coefficient. Figure 1a shows two synthetic trees with a poor match with differing locations, heights and overall crown size (Jaccard 0.01). Figure 1b shows an improved commission for location and crown size; however, some commissioning differences remain (Jaccard 0.25). Figure 1c shows a close alignment in size and location, with small commission losses in height, resulting in a close match (Jaccard 0.9), whilst Figure 1d shows a low commission between height, crown size and location (Jaccard 0.15). Figure 1e shows a close match in location, but a low match in crown height and size (Jaccard 0.40) and Figure 1f shows an offset in the location, similar crown size and minor differences in height (Jaccard 0.74).

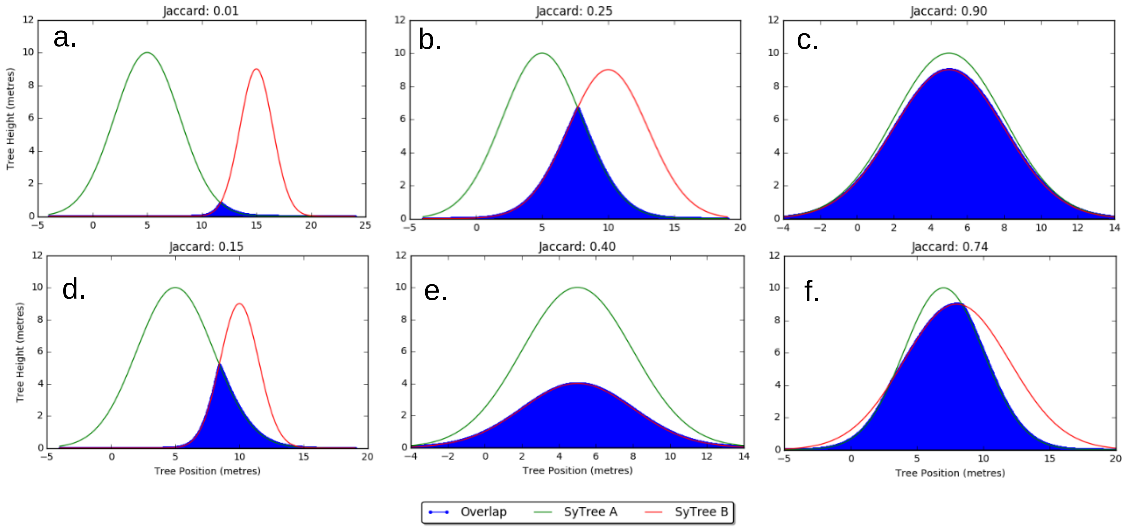


Figure 1 **Gaussian overlap used for measuring data agreement between two data sets, where the difference between the two shapes is quantified using the Jaccard similarity coefficient.**

3.3 Optimal Algorithm for Matching Populations of Trees Represented in both RS-derived and Ground Reference Datasets

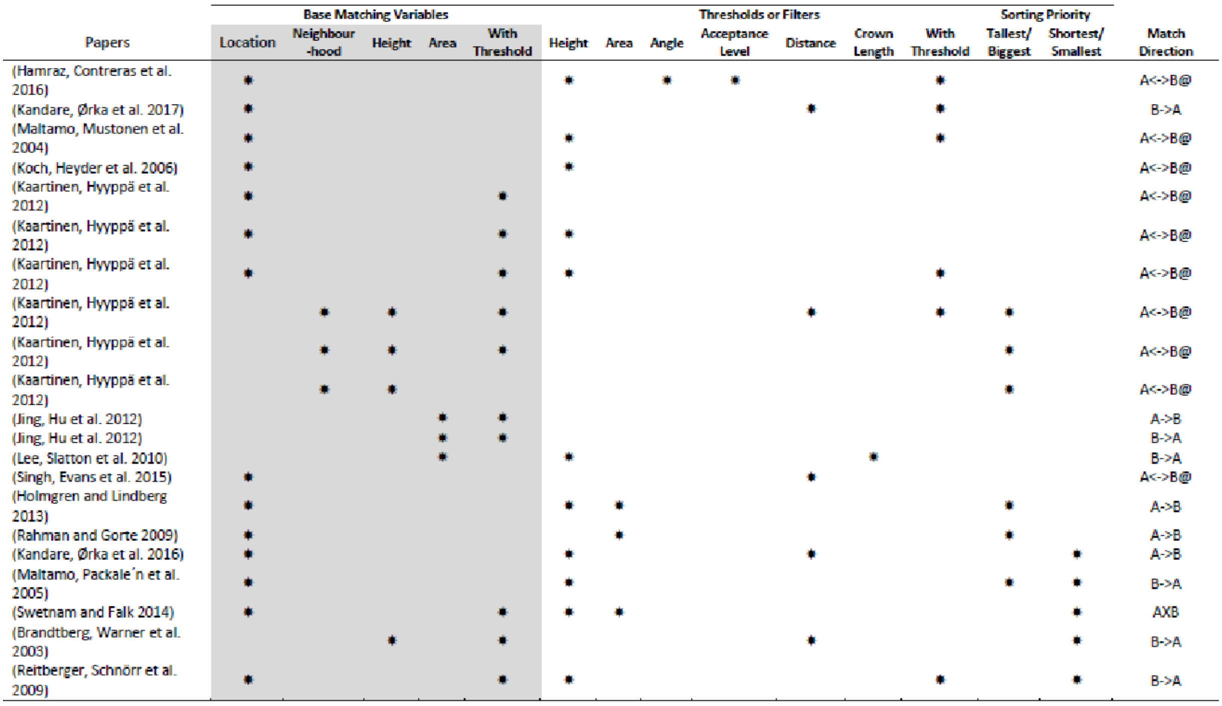
3.3.1 Meta-study of Alternative Match-pairing Methods

Following a review of highly-cited papers from peer-reviewed journals, published 2003-2017, it is apparent that many different match-pairing methods are used when evaluating agreement between GR and RS-derived data. These match-pairing methods have been consolidated into Table 1, where similar methods are grouped together (base matching method, filtered or thresholded, and sorting priority). These groups are further subdivided into methodological categories including, for example; data filtering by height, area, distance and angle. Table 1 also shows where a threshold has been applied either to the base or secondary matching filters. The direction of the match for each method is indicated as; 1) matching the GR to the RS-derived data, 2) matching RS-derived to the GR data, or 3) attempting a match in one direction, then in the other (bidirectionality) and selecting the match with the highest agreement. All of these different matching directions can potentially lead to different pairs of trees being matched, across the varying permutations. Following the review (Table 1), two representative-match-pairing (RMP) methods are defined, that replicate common match-pairing methods used in the literature:

- **RMP 1: Hausdorff Distance Algorithm**
  (Trees paired by distance to one another, the closest achieving a pair)
- **RMP 2: Within Neighbourhood, Sorted by Area and within a Height Threshold**
  (Sort A by area. Define neighbourhood of 21m. Find trees within 5m of one another, and closest sized crown areas are matched)

These two RMP methods were subsequently compared to a new approach (see 3.3.2 Hungarian Combinatorial Optimisation Algorithm) in a test using synthetic tree data (3.4 Testing the Pairwise Matching Algorithms with Synthetic Data).

Table 1 A meta-study of several match-pairing methods showing the base matching method, and identifying whether subsequent filters or thresholds are applied. The direction of the match is also shown.



	Base Matching Variables					Thresholds or Filters							Sorting Priority		
Papers	Location	Neighbour-hood	Height	Area	With Threshold	Height	Area	Angle	Acceptance Level	Distance	Crown Length	With Threshold	Tallest/ Biggest	Shortest/ Smallest	Match Direction
(Hamraz, Contreras et al. 2016)	●					●		●	●			●			A<->B@
(Kandare, Ørka et al. 2017)	●									●		●			B->A
(Maltamo, Mustonen et al. 2004)	●					●						●			A<->B@
(Koch, Heyder et al. 2006)	●					●									A<->B@
(Kaartinen, Hyyppä et al. 2012)	●				●										A<->B@
(Kaartinen, Hyyppä et al. 2012)	●				●	●									A<->B@
(Kaartinen, Hyyppä et al. 2012)	●				●	●						●			A<->B@
(Kaartinen, Hyyppä et al. 2012)		●	●		●					●		●	●		A<->B@
(Kaartinen, Hyyppä et al. 2012)		●	●		●								●		A<->B@
(Kaartinen, Hyyppä et al. 2012)		●	●										●		A<->B@
(Jing, Hu et al. 2012)				●	●										A->B
(Jing, Hu et al. 2012)				●	●										B->A
(Lee, Slatton et al. 2010)				●		●					●				B->A
(Singh, Evans et al. 2015)	●									●					A<->B@
(Holmgren and Lindberg 2013)	●					●	●						●		A->B
(Rahman and Gorte 2009)	●						●						●		A->B
(Kandare, Ørka et al. 2016)	●					●				●				●	A->B
(Maltamo, Packale´n et al. 2005)	●					●							●	●	B->A
(Swetnam and Falk 2014)	●				●	●	●							●	AXB
(Brandtberg, Warner et al. 2003)			●		●					●				●	B->A
(Reitberger, Schnörr et al. 2009)	●				●	●						●		●	B->A

Notes: A = Ground reference (GR) data. B = RS-derived (RS) data. A->B = GR matched on to RS. B->A = RS matched on to GR. A<->B@ = match attempted in both directions and the best match chosen. AXB = match directionality not described.

3.3.2 Hungarian Combinatorial Optimisation Algorithm

The Hungarian algorithm (also called the Kuhn–Munkres algorithm or Munkres assignment algorithm) is described in detail by Kuhn (1955). The Hungarian algorithm was originally defined to resolve the “assignment problem” in operations mathematics (Kuhn 1955), and has been used widely in data science, but rarely in RS or environmental studies. In this approach, the description of the data size and suitability of a match available is used in the algorithm, meaning the biophysical properties of trees from each dataset; location, height and crown area are also analysed, thereby meeting the metric criteria. The Hungarian algorithm attempts all possible pairing combinations for each point in data A against each point in data B and then *vice-versa* and outputs the optimal overall match-pairing.

3.3.3 Quantification of Accuracy with which Delineations Estimate Biophysical Properties and Population Size

Following the completion of match-pairing and Gaussian overlap assessment two accuracy metrics were calculated. The match-pairing success is quantified by the average match-pairing similarity index (AMPS). This function is the average match-pairing agreement as measured using the Gaussian overlap method (3.2 Gaussian Overlapping and the Jaccard Similarity Coefficient) calculated across all tree pairings. Higher AMPS values indicate a better overall quality of match for the paired trees. In addition to AMPS, the relative dataset sizes are also quantified to identify disparities in tree population size in GR and RS-derived datasets, for example, to show the effects of pairing directionality. The dataset size similarity index (DSS) is defined as the comparison between the total number of trees in the two datasets A and B, against the number of match-pairings achieved, expressed as a normalised value. As with AMPS, high DSS scores are preferred as this indicates similar tree population sizes in the two datasets.

3.4 Testing the Pairwise Matching Algorithms with Synthetic Data

3.4.1 Synthetic Data Environment

A synthetic environment was created to compare the biophysical attributes of RS trees, using common tree structure values typically output from ITC delineation. For simplicity, the synthetic tree (^{sy}Tree) attributes used were a known location, a predefined crown shape (circle), and a known crown area. During initial testing a single tree was modelled, ^{sy}Tree A, where the biophysical attributes of a real-world tree was randomly selected from within the 5th to 95th percentile of a broadleaved GR tree sample. By taking the biophysical attributes of ^{sy}Tree A, and using randomised offsetting of ^{sy}Tree A's location, changing the height and crown area values, a second tree was created, ^{sy}Tree B. The biophysical attribute alterations were recorded as 'known changes' between the two ^{sy}Tree populations. In subsequent testing phases, similar to the work of Romanczyk, van Aardt et al. (2013), a synthetic environment was used to simulate a complex woodland area containing 500 new ^{sy}Trees (^{sy}Tree A_{500}). As before, the ^{sy}Tree A_{500} population was subject to randomised location, height and crown area changes, further creating a secondary population, ^{sy}Tree B_{500}. This produced trees ranging from 3 to 14m tall, with crown diameters between 0.75 and 1.4 times the size of the sampled GR tree average. This procedure ensured that all 500 ^{sy}Trees had intra- and inter-population biophysical attribute differences. The recorded alterations were used as a known changes index for measuring predicted differences between ^{sy}Tree A_{500} and ^{sy}Tree $B_{500,}$ against the observed differences. Variation from the known changes index identified commission error. Figure 2 depicts 500 ^{sy}Trees, showing a) tree canopies in the predicted reference phase, and

b) following data noise and population losses. The ^{sy}Tree crowns are organised by height, replicating the presentation of the data as though observed in a CHM.

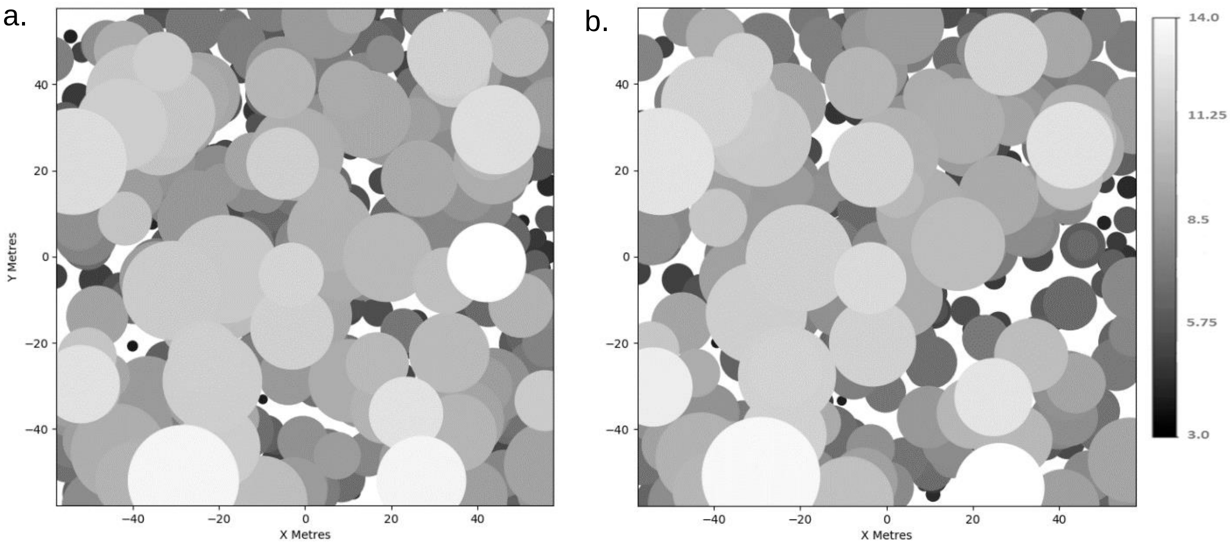


Figure 2 **500 synthetic trees representing ground reference (GR), and RS-derived LiDAR datasets. a) models 500 GR trees, and b) represents RS-derived trees with increased noise and tree losses. This replicates typically observed effects in aerial LiDAR derived canopy height models.**

3.4.2 Introduced Data Noise and Population Losses

Sensitivity testing between the ^{sy}Tree populations was undertaken by increasing data noise levels and population losses, to intentionally imbalance the datasets. The ^{sy}Tree A population remained unchanged while the ^{sy}Tree B population received randomised changes in location, height and crown area on an incremental scale (1-5). Each randomised variable used an individual set of Gaussian curves replicating the common commission problems that occur between RS-derived and GR datasets. Figure 3 illustrates changes in the location variable as each biophysical parameter had a unique set of curves. The biophysical properties of the ^{sy}Tree B population were modified by +/- of a random sample, within the appropriate distribution, relative to the prescribed noise level (Table 2). Data population losses were simulated by removing a randomised amount in incremental steps of 10% of the dataset up to a maximum of 50% removal. The introduction of data noise and loss from the tree populations, was applied across all iterations of match-pairing algorithms, to test the robustness of the different pairing methods.

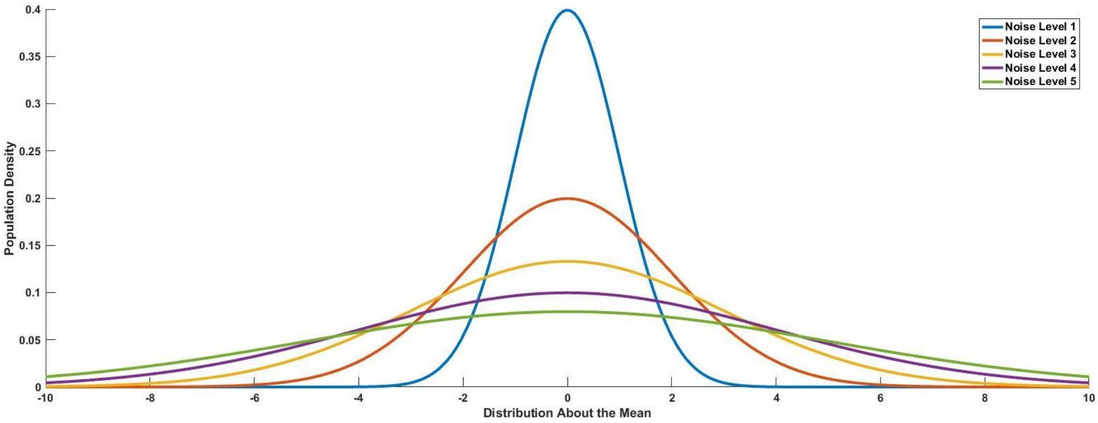


Figure 3 **An example of Gaussian curves demonstrating the change on data distribution and population density for synthetic tree data. This example represents the change in location data with the x-axis equating to metres offset. This method intentionally introduces data noise to a remote sensing dataset of synthetic trees.**

Table 2 **Introduction of data noise following modification of the normal distribution and standard deviation (SD) effect on the data population relative to data noise levels.**

Data Noise Level	Population (%) by Standard Deviation (SD)
1	SD1 = 68% +/-1, 95% +/-2, 99% +/-3
2	SD2 = 68% +/-2, 95% +/-4, 99% +/-6
3	SD3 = 68% +/-3, 95% +/-6, 99% +/-9
4	SD4 = 68% +/-4, 95% +/-8, 99% +/-12
5	SD5 = 68% +/-5, 95% +/-10, 99% +/-15

3.4.3 Results of Pairwise Matching Tests

To measure the tolerance between the predicted reference (dataset A) and observed values (dataset B), normalised root mean squared error (NRMSE) was calculated for each match-pairing method; RMP1 (Hausdorff distance), RMP2 (neighbourhood and area), and a new method, Hungarian with Gaussian overlap (Figure 4a-f). NRMSE describes the distance of the residuals from the predicted 1:1 line on a normalised scale (Figure 4a-c). This quantifies the match-pairing performance against the expected known changes index. Low NRMSE scores are preferable to high scores, hence within Figure 4a-c the scale bar is inverted. Each match-pairing method was tested with incremental data noise (level 0-5), and data population losses (0-50%). A ratio of matched-pairs was calculated for each data population (Figure 4d-f). For example, if 50 trees from 500 is paired, this achieves a paired ratio of 0.1, while pairing 450 trees achieves a paired ratio of 0.9.

Figure 4a establishes that RMP1, the Hausdorff distance match-pairing method, at noise level 0.25, achieves ~0.6 NRMSE. Furthermore, a small increase in the noise level to 0.5, significantly reduces the efficacy of the RMP1 method in achieving match-pairing to ~1.0 NRMSE. This is a uniform response across all additional levels of noise and all combinations of data population losses. In Figure 4d, the paired achieved measure for RMP1, shows a paired ratio score of 1.0 across all combinations of noise and loss. This unidirectional method demonstrates a complete data population pairing between the A and B datasets, where the matching is completed in the direction of B-A.

Figure 4b & e shows the RMP2 match-pairing method (neighbourhood and area). In comparison to Figure 4a & d, there is an uplift in results, with ~0.0 NRMSE achieved at 0 noise and 0% loss. Within Figure 4b the NRMSE score is maintained across the same level of data noise. However, a gradual increase in data noise up to level 1 rapidly diminished the NRMSE to ~0.6, at the 0% loss level. The trend follows throughout that as noise and loss increases, the NRMSE results indicate a worsening match-pairing performance. This continues to noise level 1.5, where the NRMSE values across all amounts of data loss are between ~0.9 to ~1.0 NRMSE. Figure 4e indicates that very low levels of noise is tolerated throughout all permutations of data losses (1.0 NRMSE at noise level 0). Only marginal increases in data noise, to 0.25, rapidly reduce the pairing ratio to ~0.6. At the point of noise level 1 the paring ratio has decreased to ~0.1 across all permutations. At noise level 2, the pairing ratio is reduced to 0.0. Figure 4e demonstrates this bidirectional method achieves a full pairing ratio of 1.0 across all data losses to 50% at noise level 0. A marginal increase in noise to 0.25 reduces the paired matching ratio to ~0.6 across all losses. This rapid decrease continues to noise level 1, where only a ~0.2 paired ratio is achieved, and by noise level 1.5, the paired ratio further reduces to ~0.0. Therefore, this bidirectional routine is demonstrably affected by the data losses applied.

Figure 4c and f shows the new approach of using the Hungarian and Gaussian overlap match-pairing method. Within Figure 4c this method maintains 0.0 NRMSE across all data loss levels, up to the 0.5 noise level. At noise level 1, the analysis shows a low reduction to ~0.1 NRMSE across all data loss levels to 50%, which is a significant improvement over the previous two match-pairing methods at the same noise level. There is a further increase to ~0.2 NRMSE at noise level 2, again, this is broadly spread across all loss levels. Figure 4c shows that from this noise level, the metric achieves low incremental rises in NRMSE scores, with the method achieving ~0.6 NRMSE at noise level 3. This continues up to the highest noise level of all of the match-pairing methods, where at noise level 3.75 a ~1.0 NRMSE is reached. Figure 4f identifies that throughout all combinations of increasing data noise, the Hungarian and Gaussian overlap match-pairing method maintains the ideal paired ratio 1.0, withstanding all effects of data loss up to 50%. This bidirectional, optimised method outperforms the RMP2 method in paired ratio results and equals the paired ratio output for RMP1.

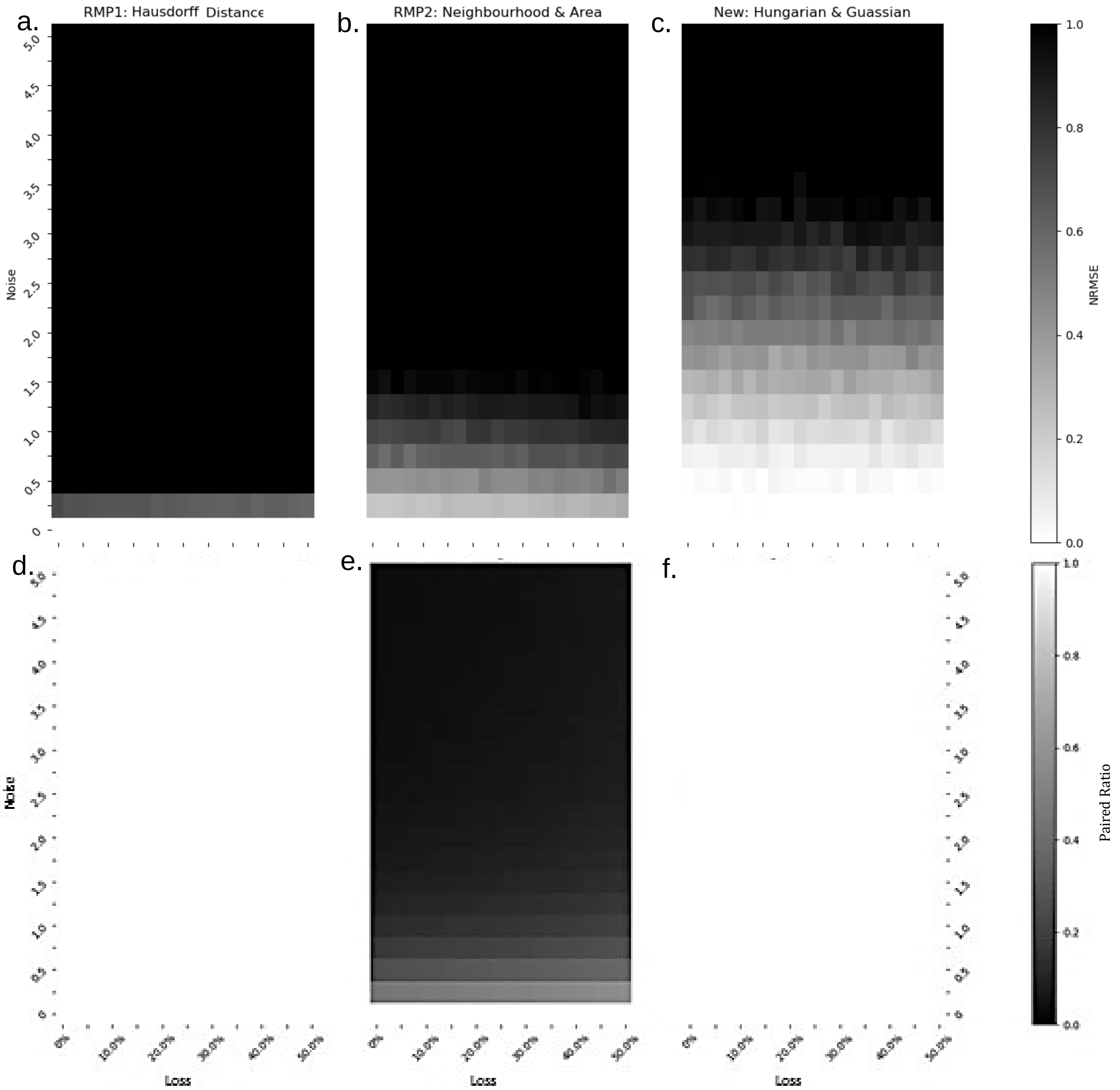


Figure 4 A combination of three data match-pairing methods being tested for the ability to achieve predicted data pairings between synthetic GR and RS-derived data. Each pixel in plots a-c represents an assessment of normalised root mean squared error (NRMSE) at differing levels of data noise and loss. Plots d-f represent the effect of the match-pairing on the data population, expressed as a pairing ratio.

3.4.4 Summary Observations and Recommendation

RMP1 (the Hausdorff distance method), for almost all of the possible data noise and loss combinations, fails to provide reliable match-pairings against the known changes. The method computes ~1.0 NRMSE from very low levels of data noise (Figure 4a). The inability to accommodate this noise is due to the way the Hausdorff algorithm uses a linear distance measure between the edges of two shapes. In this application, this is the outer edges of two ITC tree crowns. Correspondingly, the Hausdorff distance score reduces the closer the crowns are to one another, before the crown edges touch when reaching a ‘union’. The situation changes, however, at the point that the crown edges begin to intersect (Maroševič 2018). Where a smaller crown passes inside a larger crown, as is typical when aligning GR and RS-derived trees, the Hausdorff distance increases as the crown edges begin to move away from each other and the crowns wholly overlap, despite the crown centroids not yet being aligned (Marošević 2018). This makes the Hausdorff distance algorithm unreliable in match-pairing using circular crowns. In considering the data population, Figure 4d demonstrates a paired ratio of 1.0 for the unidirectional method. As the match-pairing runs, the algorithm seeks matches for all trees within the response dataset B. When all the matches in B are filled against A, the algorithm is completed and returns the ratio 1.0 (100% matched). Achieving the paired ratio of 1.0 is maintained up to the 50% data loss, despite there being up to 50% remaining unmatched trees in the A dataset. This highlights that as the method matches in a single direction, false-positive results can be reached when data size is not reported.

RMP2, the neighbourhood and area match pairing method, demonstrates an improved performance when compared to RMP1 (Figure 4b & e). However, there is a rapid reduction in the ability of this method to accurately achieve the predicted levels of match-pairing after the introduction of very low levels of data noise (Figure 4b). This is a consequence of the neighbourhood and area thresholds that limit the amount of available matches. As shown in Figure 4b, the threshold effect is compounded rapidly with increasing data noise and population loss. Notably, Figure 4e demonstrates that despite the bidirectional matching routine, the pairing ratio rapidly decreases to ~0.1, (~50 trees) at noise level 1.5. During bidirectional matching, A is matched to B, then B to A, and the best match retained (A=B). However, the implication is that the match-pairing may not necessarily occur with the same trees, for example, A matches to B, but B matches to a third tree (B=C), therefore A≠B, so A is discarded without a match. This effect, and the influence of up to 50% data losses, means that the bidirectional, RMP2 method, artificially reports acceptable levels of matches only with the reduced numbers of trees that remain. Significantly, the number of true matches achieved, as demonstrated by the paired ratio is very low (Figure 4e).

The new Hungarian and Gaussian overlap match-pairing method provides the highest levels of agreement with the predicted measures, including into the highest levels of data noise (Figure 4c). The final NRMSE values are measured at more than twice the noise level achieved than RMP2. RMP1 reduced to ~1.0 NRMSE at noise level 0.5, while RMP2 achieved ~1.0 NRMSE at noise level 1.5. However, the Hungarian and Gaussian match-pairing method continues to achieve ~0.6 NRMSE at noise level 3, and finally reaching ~1.0 NRMSE at noise level 3.75. This indicates that at more than double the noise level of the next best performing method, the Hungarian and Gaussian method is considerably more robust to the influence of improper matches. The stability of this method is further demonstrated in Figure 4f, where the match-pairing method returns a paired ratio of 1.0 across all levels of data noise, and data losses. This is due to the optimised, bidirectional nature of the Hungarian algorithm. The algorithm attempts to pair all possible combinations of each data point in A, with all possible combinations of points in B, then similar to the bidirectional approach, the process is repeated *visa-versa*. However, in the Hungarian algorithm, the routine searches for a match-pair from the opposing dataset for every individual data point within the primary data, considering every possible data point in the opposing dataset, and attempting all possible parameter combinations before the best match is achieved. Therefore, this method achieves a true-positive match from all available options, and a 1.0 paired ratio score for the entire data population.

In summary, within the analysis framework conducted in a synthetic environment, the Hungarian and Gaussian curve match-pairing is demonstrated as being the most effective in accurately resolving the match-pairing problem between GR and RS-derived data. Therefore, following the metrics development and analysis phase, the Hungarian and Gaussian curve match-pairing method is the recommended approach for use in quantifying match-pairing agreement with real-world data.

3.5 The ARBOR Framework

Following the findings of the analysis and results above, the final implementation of the ARBOR framework is illustrated at Figure 5. This structure defines the developmental phase output with a simple, worked example of how the AROBR framework would interact with two datasets representing a sample of GR trees (*n*=100), and RS-derived trees for the same area (*n*=60).

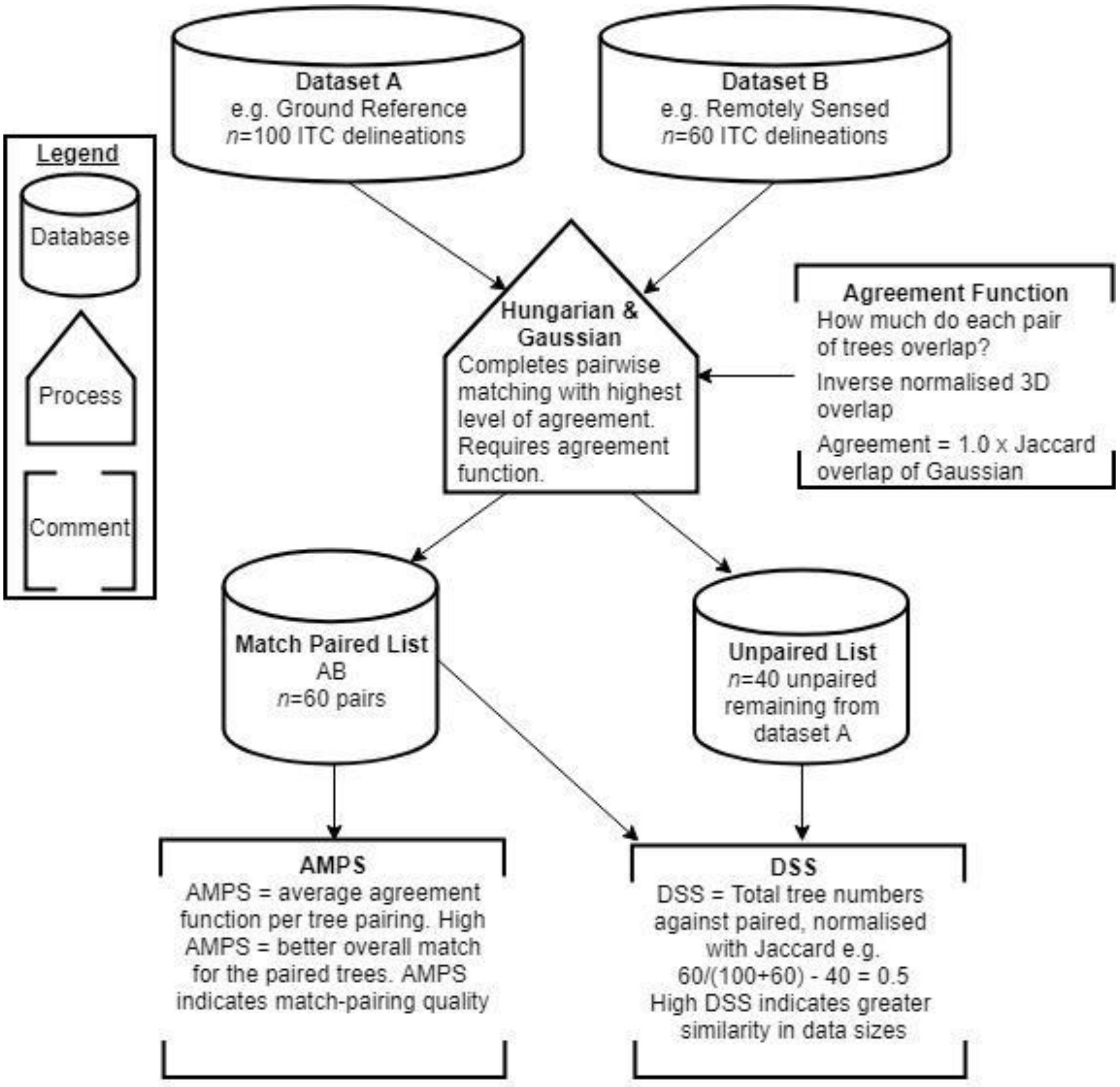

Legend
Database
Process
Comment
Dataset A
e.g. Ground Reference
n=100 ITC delineations
Dataset B
e.g. Remotely Sensed
n=60 ITC delineations
Hungarian & Gaussian
Completes pairwise matching with highest level of agreement. Requires agreement function.
Agreement Function
How much do each pair of trees overlap?
Inverse normalised 3D overlap
Agreement = 1.0 x Jaccard overlap of Gaussian
Match Paired List
AB
n=60 pairs
Unpaired List
n=40 unpaired remaining from dataset A
AMPS
AMPS = average agreement function per tree pairing. High AMPS = better overall match for the paired trees. AMPS indicates match-pairing quality
DSS
DSS = Total tree numbers against paired, normalised with Jaccard e.g.
60/(100+60) - 40 = 0.5
High DSS indicates greater similarity in data sizes

Figure 5 **A working example of the ARBOR framework workflow for the quantification of match-pairing agreement between remote sensing derived and ground reference data. Notes: AMPS = averaged matched-pairing similarity index, DSS = dataset size similarity index**

3.6 Demonstration of ARBOR for Evaluating ITC Delineations

To demonstrate the principal of the ARBOR framework for quantifying agreement between GR and RS-derived data, the model described in Figure 5, was applied to a large, broadleaved woodland study site that had been scanned by a fixed-wing aircraft, generating ALS LiDAR and digital photography data, and contained twenty-six, 20x20m GR plots, that were manually surveyed with biophysical tree attributes measured and recorded (see *supplementary information*).

The GR plots were identified in the LiDAR data and CHMs for each GR plot was created. Each GR plot was delineated using four different methods. A technician experienced in both manual tree surveying and remote sensing undertook manual ITC delineation (ITC$_{MAN}$) by digitising vector polygons in ESRI ArcGIS, using a similar approach as described in Brandtberg and

Walter (1998). The polygon followed tree crown edges on the CHM, defining crown outlines, crown areas and location centroids. Inverse watershed ITC delineation (ITC_{IWD}) is a frequently used technique (Kwak, Lee et al. 2007, Jing, Hu et al. 2014). ITC_{IWD} identifies valleys (gulleys), and in a top-down approach, locates tree crowns edges where adjacent tree crowns meet. This delineation procedure produces a network of connected valleys with the ITC_{IWD} delineated crowns as 'islands' between the valleys, and outputs a vector-defined crown edge, location and crown area (Kwak, Lee et al. 2007, Jing, Hu et al. 2014). A variable limit local maxima ITC delineation algorithm, incorporating metabolic scaling theory (MST) predictions to remove data noise (ITC_{MST}), was also used (Swetnam and Falk 2014). The ITC_{MST} method initially uses inverse watershedding delineation, but refines tree locations and assignment with MST, outputting individual tree locations, crown areas, and tree heights. Finally, a photogrammetric ITC delineation technique (ITC_{PHO}) was applied to high resolution optical imagery to define tree crown boundaries and locations. For all ITC delineation methods the resulting vector polygons provide tree crown location, centralised height points, and circular shaped tree crowns.

3.6.1 The Results of Applying ARBOR to RS-derived ITC Delineations

The delineation techniques ITC_{MAN}, ITC_{IWD}, ITC_{MST} and ITC_{PHO} were individually analysed against the GR data using the ARBOR framework, where Gaussian overlap replicates the biophysical characteristics of trees and defines the AMPS (averaged match-pairing similarity index) and DSS (dataset size similarity index) to optimise pairwise matching and to measure data population correspondence. Figure 6 demonstrates that the four ITC delineation techniques achieved varying levels of match-pairing agreement.

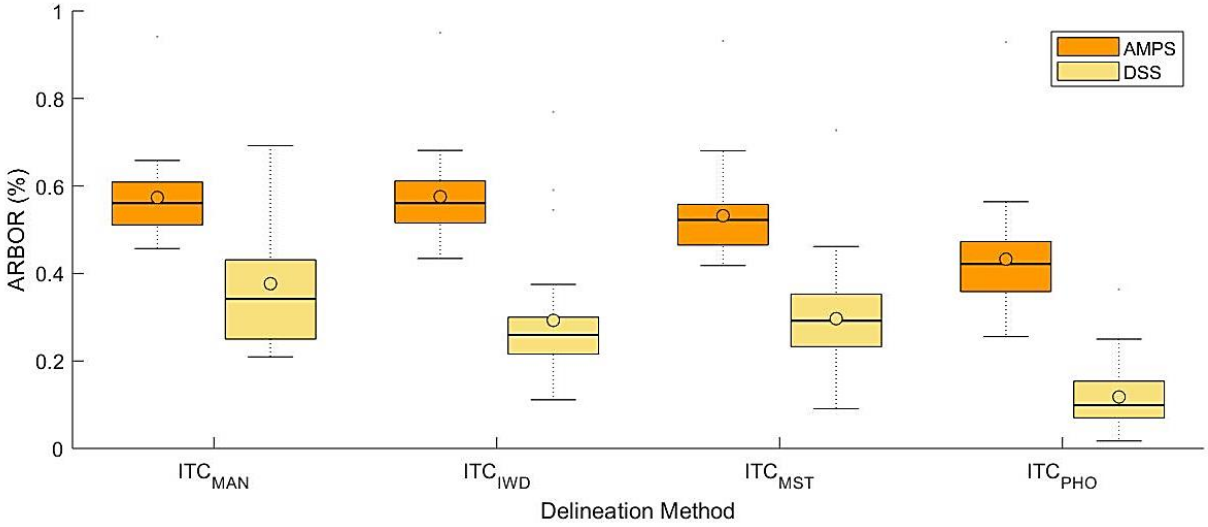


Figure 6 **ARBOR scores comparing the match-pairing success between four different ITC delineation techniques acquired from aerial LiDAR data with ground reference data over 26 survey plots.**

ITC$_{MAN}$ and ITC$_{IWD}$ have the highest AMPS values, indicating that these delineation techniques have a similar level of accuracy (Table 3). The ITC$_{MST}$ delineation also achieved a level of accuracy commensurate with the ITC$_{MAN}$ and ITC$_{IWD}$ methods, although this was marginally lower. The interquartile range (IQR) of the AMPS is similar for all four ITC methods. All four methods show marginal positive skewing in the AMPS values indicating a majority of results are to the upper end of the IQR, and that the median result is closely aligned to the first quartile (1Q) results.

The ITC$_{MAN}$ achieved the highest DSS values indicating the highest overall level of accuracy in measuring biophysical tree attributes. For the automated delineation techniques, ITC$_{IWD,}$ ITC$_{MST}$ and ITC$_{PHO}$ achieved lower DSS values of 0.26, 0.29 and 0.1 at the median respectively. The ITC$_{MAN}$ indicates a large Q3 range to the maximum (~10%). Overall, ITC$_{IWD,}$ ITC$_{MST}$ and ITC$_{PHO}$ show largely balanced distributions in their respective DSS IQR. The ITC$_{PHO}$ achieved the lowest overall ARBOR scores in both AMPS and DSS, when compared against the other delineation techniques.

In all of the results for both AMPS and DSS values across all four delineation techniques show the mean, visualised as a circle, is greater than the median line (Figure 6). This indicates there is a longer upper tail, showing a positive skew to these results. This also shows that the median result is closely aligned to the 1Q. The only exception is the DSS mean for the ITC$_{MST}$ where both the mean and median are closely aligned (Figure 6).

Table 3 **Quantification of ARBOR framework scores for four individual tree crown (ITC) delineation techniques, when compared to known tree location, height and crown areas of ground reference tree data.**

	ARBOR Framework (%)											
	AMPS						DSS					
Delineation	**Q1**	**Med**	**Mean**	**Q3**	**Min**	**Max**	**Q1**	**Med**	**Mean**	**Q3**	**Min**	**Max**
ITC$_{MAN}$	0.51	0.56	0.57	0.61	0.46	0.66	0.25	0.34	0.38	0.43	0.21	0.69
ITC$_{IWD}$	0.52	0.56	0.58	0.61	0.43	0.68	0.22	0.26	0.29	0.30	0.11	0.38
ITC$_{MST}$	0.46	0.52	0.53	0.56	0.42	0.68	0.23	0.29	0.30	0.35	0.09	0.46
ITC$_{PHO}$	0.36	0.42	0.43	0.47	0.26	0.56	0.07	0.10	0.12	0.15	0.02	0.25

Notes: AMPS = averaged matched-pairing similarity index, DSS = dataset size similarity index, MAN = manual, IWD = inverse watershedding, MST = variable limit maxima with metabolic scaling theory, PHO = photogrammetric method.

The application of ARBOR to RS-derived ITC delineation and GR data, demonstrates how the framework can quantify differences in ITC delineation techniques, and allows a discriminatory assessment for identifying the ITC delineation technique which would achieve the highest levels of accuracy for the data user.

4.0 The Significance of the ARBOR Framework

Culvenor (2002) states that achieving the successful delineation of trees is problematic. Outlining trees from homogenous groups, without explicitly quantified GR data can lead to repeated errors. The aim of this study was to develop a framework for objectively quantifying the agreement between two datasets, focussing on common commission errors in RS data, with increased data noise and data population differences. The ARBOR framework was developed and then applied to real-world data to quantify the commission agreement between four different ITC delineation techniques and GR datasets (Figure 6). This type of analysis is frequently absent from RS studies that utilise ITC delineation techniques, which instead, rely upon arbitrary height or other cut-off thresholds to infer the level of agreement (Næsset 2002, Listopad, Drake et al. 2011, Hyyppa, Yu et al. 2012). However, the findings from this research indicates that simple measures, thresholding and not accounting for the biophysical parameters of trees leads to low levels of true-positive match-pairing between GR and RS-derived data (Figure 4).

Throughout Figure 4a-f, there is a general tendency of higher match-pairing performance at lower noise levels, with a diminishing of NRMSE as noise levels increase. Concurrently, increasing data loss, from 0 to 50%, further impacts on the efficacy of the match-pairing. In all cases, noise affecting the data has the greatest effect, while data loss, less so. What is clear is that introducing data noise alters the biophysical parameters that the trees are being matched on, and therefore, assessment of these parameters should always be included as variables when seeking ITC delineation agreement with GR data. Figure 4a-c shows that match-pairing methods are sensitive to shifts in the biophysical tree structure under analysis. The data losses, or differences in tree population numbers between the two datasets, has a different effect. Where data in the observed dataset B (e.g. LiDAR) has fewer trees, poorer matches are achieved as the limited tree population will have greater tree numbers available for matching in the opposing dataset A (e.g. GR). Using some methods, such as Hausdorff distance, unmatched tree data is discarded from the analysis when all trees in dataset B are matched. Without measuring the dataset size, the match-pairing analysis declares a successful match even where there are fewer trees in one set than the other. This creates a false positive result, where changes in the data population and quantification of the unmatched pairings is not reported (Figure 4d-e). Furthermore, this analysis has shown that the frequently used match-pairing method, Hausdorff distance, significantly underperforms in reaching agreement between GR and RS datasets, particularly when exposed to increasing data noise and losses, as readily occurs in real-world RS data (Figure 4a & d). However, through the creation of the ARBOR framework, a demonstrably robust framework has been established to quantify agreement between GR and RS-derived data.

The approach used to develop the ARBOR framework was similar to Ørka, Næsset et al. (2009), where a synthetic testing environment was used to replicate complex RS tree datasets, with naturally occurring variations in tree size, shape and location. During early iterations of metric testing, it was recognised that each tree in the two datasets must achieve a bilateral matching agreement. However, this was problematic as it was observed that this lead to 'hugging pairs' within the data assignment. Specifically, where once assigned a matched pair, e.g. ^{SY}Tree A1 to ^{SY}Tree B1, the assignment excluded any other potential match even where a subsequent potential match was better suited. Further analysis showed that the order of the match-agreement process is a relevant factor in achieving high agreement match-pairing. To overcome this problem, the Hungarian combinatorial optimisation algorithm was used to search through all the potential combinations in the parallel dataset. An advantage of the Hungarian algorithm is the optimising nature of the routine where the algorithm cannot reach completion with an unsuitable data assignment. Therefore, the algorithm attempts all possible data combinations between the two datasets and completes only when the fullest level of agreement is reached.

The AMPS index quantifies the similarity between the datasets as a measure of the biophysical tree properties agreement, represented as Gaussian overlap (Figure 1), while the DSS index provides a measure of population size estimates from ITC delineations. Contrary to the views of Kaartinen, Hyyppä et al. (2012), who state that the comparison of delineation results between different datasets cannot be achieved due to the variability in crown structures of different species, this research demonstrates that by using GR representations of trees as simple objects (with location, height and area), and matching these objects to ITC delineations using a Gaussian curve model and the Hungarian algorithm, accuracy assessment becomes possible (Figure 6). Therefore, the ARBOR framework provides a new opportunity for quantifying the confidence of ITC delineation techniques in RS investigations. Figure 6 and Table 3 demonstrate that recommendations can be given about the efficacy and suitability of different ITC delineation techniques applied to remotely-sensed data. We can define optimal ITC delineation methods, as shown by the AMPS and DSS values calculated within the ARBOR framework.

In Figure 6 the AMPS and DSS scores appear to be low for all delineation techniques, given that they could potentially rise to a value of 1 in the case of perfect matches. In order to explain the low scores shown in Figure 6, it is worth noting that our reference data was collected in the field and all trees >5cm DBH were recorded, meaning that many trees may have been understorey trees or not exposed as full crowns at the top of the forest canopy. Hence, the low DSS scores are likely to represent the large number of understory trees shadowed by more dominant trees and therefore not clearly defined in the LiDAR data. Low AMPS scores reflect the differences in biophysical properties as expressed in GR and ITC delineations and this may be explained in part by the errors in both field and ITC delineation methods, as discussed previously. For example, it is well recognised that penetration of LiDAR signals into the tree canopy can result in an underestimation of tree height, which may be inconsistent between tree of differing species and crown characteristics (Næsset, 1997). Furthermore, trees exhibit a natural structural variance which Mandelbrot (1982) notes is sculpted by ‘chance, irregularities and non-uniformity’. Low AMPS scores are reflective of the natural complexities that are observed in tree crown structure, which may be difficult to detect in the simplified descriptions of crown geometry in both field and ITC delineation data.

When matching reference data to ITC delineations there can be data disparities in both directions, e.g. several small adjacent trees can be delineated as one large tree in the ITC and *vice versa*. ARBOR matches trees in both directions, from reference to ITC delineation and again in the opposite direction. This approach means that a quantification of the errors can be made in the examples highlighted above. Where there is a lack of matching it follows that there are lower AMPS and DSS scores. For example, where 1 large whole tree in the reference data is matched to an incorrectly identified tree in the ITC delineation data which is actually only a subcomponent of the large tree canopy, the AMPS score will be lower due to poor correspondence in the biophysical properties of the matched trees. As another example, where many smaller trees in the reference data have been erroneously identified as one large tree in the ITC delineation, only one of the small trees will be matched to the ITC data; this will depress the DSS score due to the numbers of trees in each dataset being poorly matched. The ARBOR tool can be used to isolate individual occurrences of mis-agreement between reference and ITC delineations. This allows a user to investigate the reasons for this mis-agreement and implement appropriate improvements in the ITC delineation procedure.

The principal emphasis of this work was to enable the quantification of pairwise match agreement between GR and RS-derived datasets. However, we also recognise there are opportunities for the ARBOR framework to quantify other types of data agreement, for example, tree delineations derived from aerial photography matched with those from aerial or terrestrial LiDAR. Due to the modular nature of the ARBOR framework, it can be adapted, as is required in future studies, to include a range of different match-pairing metrics not incorporated into this study and to generate alternative statistical measures of ITC delineation accuracy. Furthermore, in this study the ARBOR framework was used for quantifying the accuracy of ITC delineation in a complex semi-natural temperate broadleaved woodland. Given the demonstrable robustness of the tree matching technique and sensitivity of the accuracy metrics, the ARBOR framework holds potential as an objective and transferable tool that can be applied across the full range of forest types.

To enable the distribution and further application of the ARBOR framework, a portal has been developed to allow the uploading and analysis of match-pairing data, to provide objective quantification of the accuracy of ITC delineations. <<<*NOTE for Editor/reviewers: a fully functioning site with a flexible user interface will be up and running at the time of this paper being published and the URL will be inserted at this point in the manuscript* >>>

5.0 Conclusion

It is recognised that achieving accurate ITC delineation is a difficult task, particularly in broadleaved tree crowns. Currently there are no standardised techniques or measures of the amount of agreement between RS-derived and GR datasets. Many potential errors arise in the alignments of these data, however, a common approach to addressing these errors is to apply arbitrary cut-off thresholds. These thresholds are intended to determine whether the same individual tree is identified within the two different datasets, but there are limitations in these approaches, particularly as some match-pairing methods can lead to false-positive results. Furthermore, the reporting of ITC delineation accuracy is limited in general. Through the use of a synthetic test environment, an optimised algorithm was identified for matching RS-derived and GR tree populations and statistical metrics were developed for quantifying ITC delineation accuracy based on biophysical attributes and data population size. These methods were incorporated into the ARBOR framework which provides a practical approach for achieving and quantifying match-pairing agreement between RS-derived and GR datasets. Therefore, the ARBOR framework is proposed as a standardised solution for future ITC delineation accuracy assessment.

6.0 Supplementary Information

Supplementary information is included with this submission.

7.0 Acknowledgements

The authors would like to thank the anonymous reviewer for their helpful comments in improving this paper, and would like to thank NERC ARF for their contribution to this research through the provision of facilities and resources for the capture of the remotely-sensed data of the study site. This research was supported by an EPSRC studentship for the lead author: EP/L504804/1.

8.0 References

Brandtberg, T. and F. Walter (1998). "Automated delineation of individual tree crowns in high spatial resolution aerial images by multiple-scale analysis." Machine Vision and Applications **11**(2): 64-73.

Culvenor, D. S. (2002). "TIDA: an algorithm for the delineation of tree crowns in high spatial resolution remotely sensed imagery." Computers & Geosciences **28**(1): 33-44.

Duncanson, L. I., R. O. Dubayah, B. D. Cook, J. Rosette and G. Parker (2015). "The importance of spatial detail: Assessing the utility of individual crown information and scaling approaches for lidar-based biomass density estimation." Remote Sensing of Environment **168**: 102-112.

Eysn, L., M. Hollaus, K. Schadauer and N. Pfeifer (2012). "Forest Delineation Based on Airborne LIDAR Data." Remote Sensing **4**(3).

Hladik, C. and M. Alber (2012). "Accuracy assessment and correction of a LIDAR-derived salt marsh digital elevation model." Remote Sensing of Environment **121**: 224-235.

Holmgren, J. and E. Lindberg (2013). "Tree Crown Segmentation Based on a Geometric Tree Crown Model for Prediction of Forest Variables." Canadian Journal of Remote Sensing **39**(S1): S86-S98.

Hyyppa, J., X. W. Yu, H. Hyyppa, M. Vastaranta, M. Holopainen, A. Kukko, H. Kaartinen, A. Jaakkola, M. Vaaja, J. Koskinen and P. Alho (2012). "Advances in Forest Inventory Using Airborne Laser Scanning." Remote Sensing **4**(5): 1190-1207.

Jakubowksi, M. K., Q. Guo, B. Collins, S. Stephens and M. Kelly (2013). "Predicting Surface Fuel Models and Fuel Metrics Using Lidar and CIR Imagery in a Dense, Mountainous Forest." Photogrammetric Engineering & Remote Sensing **79**(1): 37-49.

Jing, L., B. Hu, H. Li, J. Li and T. Noland (2014). Automated individual tree crown delineation from LIDAR data using morphological techniques. 35th International Symposium on Remote Sensing of Environment. H. Guo. Bristol, Iop Publishing Ltd. **17**.

Jing, L., B. Hu, T. Noland and J. Li (2012). "An individual tree crown delineation method based on multi-scale segmentation of imagery." ISPRS Journal of Photogrammetry and Remote Sensing **70**: 88-98.

Kaartinen, H., J. Hyyppä, X. Yu, M. Vastaranta, H. Hyyppä, A. Kukko, M. Holopainen, C. Heipke, M. Hirschmugl, F. Morsdorf, E. Næsset, J. Pitkänen, S. Popescu, S. Solberg, B. M. Wolf and J.-C. Wu (2012). "An International Comparison of Individual Tree Detection and Extraction Using Airborne Laser Scanning." Remote Sensing **4**(4): 950.

Kandare, K., H. O. Ørka, J. C.-W. Chan and M. Dalponte (2016). "Effects of forest structure and airborne laser scanning point cloud density on 3D delineation of individual tree crowns." European Journal of Remote Sensing **49**(1): 337-359.

Koch, B., U. Heyder and H. Weinacker (2006). "Detection of Individual Tree Crowns in Airborne Lidar Data." Photogrammetric Engineering & Remote Sensing **72**(4): 357-363.

Kuhn, H. W. (1955). "The Hungariain Method for the Assignment Problem." Naval Research Logistics Quarterly **2**: 83-97.

Kwak, D.-A., W.-K. Lee, J.-H. Lee, G. S. Biging and P. Gong (2007). "Detection of individual trees and estimation of tree height using LiDAR data." Journal of Forest Research **12**(6): 425-434.

Leckie, D. G., N. Walsworth and F. A. Gougeon (2016). "Identifying tree crown delineation shapes and need for remediation on high resolution imagery using an evidence based approach." ISPRS Journal of Photogrammetry and Remote Sensing **114**: 206-227.

Listopad, C., J. B. Drake, R. E. Masters and J. F. Weishampel (2011). "Portable and Airborne Small Footprint LiDAR: Forest Canopy Structure Estimation of Fire Managed Plots." Remote Sensing **3**(7): 1284-1307.

Lu, X. C., Q. H. Guo, W. K. Li and J. Flanagan (2014). "A bottom-up approach to segment individual deciduous trees using leaf-off lidar point cloud data." Isprs Journal of Photogrammetry and Remote Sensing **94**: 1-12.

Mandelbrot, B. (1982). The Fractal Geometry of Nature. San Francisco: W. H. Freeman and Company.

Marošević, T. (2018). "The Hausdorff Distance Between Some Sets of Points." Mathematical Communications **23**(2): 247-257.

Nowakowska, E., J. Koronacki and S. Lipovetsky (2014). "Tractable Measure of Component Overlap for Gaussian Mixture Models." arXiv **1407**(7172.1): 1-24.

Nowakowska, E., J. Koronacki and S. Lipovetsky (2015). "Clusterability assessment for Gaussian mixture models." Applied Mathematics and Computation **256**: 591-601.

Næsset, E. (1997). Determination of mean tree height of forest stands using airborne laser scanner data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 52(2), 49-56. doi:https://doi.org/10.1016/S0924-2716(97)83000-6.

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

Ole Ørka, H., E. Næsset and O. M. Bollandsås (2009). "Classifying Species of Individual Trees by Intensity and Structure Features Derived from Airborne Laser Scanner Data " Remote Sensing of Environment **113**: 1163–1174.

Rahman, M. Z. A. and B. G. H. Gorte (2009). Tree Crown Delineation from High Resolution Airborne LiDAR Based on Densities of High Points. Laser Scanning, Paris, ISPRS.

Romanczyk, P., J. van Aardt, K. Cawse-Nicholson, D. Kelbe, J. McGlinchy and K. Krause (2013). "Assessing the impact of broadleaf tree structure on airborne full-waveform small-footprint LiDAR signals through simulation." Canadian Journal of Remote Sensing **39**(sup1): S60-S72.

Singh, M., D. Evans, B. S. Tan and C. S. Nin (2015). "Mapping and Characterizing Selected Canopy Tree Species at the Angkor World Heritage Site in Cambodia Using Aerial Data." PLOS ONE **10**(4): e0121558.

Swetnam, T. L. and D. A. Falk (2014). "Application of Metabolic Scaling Theory to reduce error in local maxima tree segmentation from aerial LiDAR." Forest Ecology and Management **323**: 158-167.

West, P. W. (2009). Tree and Forest Measurement. London: Springer Science & Business Media.

Wu, B., B. Yu, Q. Wu, Y. Huang, Z. Chen and J. Wu (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.

Yu, X., J. Hyyppä, A. Kukko, M. Maltamo and H. Kaartinen (2006). "Change Detection Techniques for Canopy Height Growth Measurements Using Airborne Laser Scanner Data." Photogrammetric Engineering & Remote Sensing **72**(12): 1339–1348.

Yu, X., J. Hyyppä, P. Litkey, H. Kaartinen, M. Vastaranta and M. Holopainen (2017). "Single-Sensor Solution to Tree Species Classification Using Multispectral Airborne Laser Scanning." Remote Sensing **9(2)**(108): 1-16.

Zhao, K., J. C. Suarez, M. Garcia, T. Hu, C. Wang and A. Londo (2018). "Utility of multitemporal lidar for forest and carbon monitoring: Tree growth, biomass dynamics, and carbon flux." Remote Sensing of Environment **204**: 883-897.

Zhen, Z., L. J. Quackenbush and L. J. Zhang (2016). "Trends in Automatic Individual Tree Crown Detection and Delineation-Evolution of LiDAR Data." Remote Sensing **8**(4): 26.

9.0 List of Figure Captions

Figure 1 **Gaussian overlap used for measuring data agreement between two data sets, where the difference between the two shapes is quantified using the Jaccard similarity coefficient.**

Figure 2 **500 synthetic trees representing ground reference (GR), and RS-derived LiDAR datasets. a) models 500 GR trees, and b) represents RS-derived trees with increased noise and tree losses. This replicates typically observed effects in aerial LiDAR derived canopy height models.**

Figure 3 **An example of Gaussian curves demonstrating the change on data distribution and population density for synthetic tree data. This example represents the change in location data with the x-axis equating to metres offset. This method intentionally introduces data noise to a remote sensing dataset of synthetic trees.**

Figure 4 **A combination of three data match-pairing methods being tested for the ability to achieve predicted data pairings between synthetic GR and RS-derived data. Each pixel in plots a-c represents an assessment of normalised root mean squared error (NRMSE) at differing levels of data noise and loss. Plots d-f represent the effect of the match-pairing on the data population, expressed as a pairing ratio.**

Figure 5 **A working example of the ARBOR framework workflow for the quantification of match-pairing agreement between remote sensing derived and ground reference data. Notes: AMPS = averaged matched-pairing similarity index, DSS = dataset size similarity index**

Figure 6 **ARBOR scores comparing the match-pairing success between four different ITC delineation techniques acquired from aerial LiDAR data with ground reference data over 26 survey plots.**