remote sensing

MDPI

Article

A Comparison of Imputation Approaches for Estimating Forest Biomass Using Landsat Time-Series and Inventory Data

Trung H. Nguyen [1,2,*], Simon Jones [1,2], Mariela Soto-Berelov [1,2], Andrew Haywood [3] and Samuel Hislop [1,2]

1 Remote Sensing Centre, School of Science, RMIT University, Melbourne, VIS 3000, Australia; simon.jones@rmit.edu.au (S.J.); mariela.soto-berelov@rmit.edu.au (M.S.-B.); samuel.hislop@rmit.edu.au (S.H.)
2 Cooperative Research Centre for Spatial Information (CRCSI), Carlton, VIC 3053, Australia
3 European Forest Institute, 08025 Barcelona, Spain; andrew.haywood@efi.int
* Correspondence: trung.nguyen3@rmit.edu.au; Tel.: +61-3-9925-2000

Received: 24 September 2018; Accepted: 15 November 2018; Published: 17 November 2018

Abstract: The prediction of forest biomass at the landscape scale can be achieved by integrating data from field plots with satellite imagery, in particular data from the Landsat archive, using k-nearest neighbour (kNN) imputation models. While studies have demonstrated different kNN imputation approaches for estimating forest biomass from remote sensing data and forest inventory plots, there is no general agreement on which approach is most appropriate for biomass estimation across large areas. In this study, we compared several imputation approaches for estimating forest biomass using Landsat time-series and inventory plot data. We evaluated 18 kNN models to impute three aboveground biomass (AGB) variables (total AGB, AGB of live trees and AGB of dead trees). These models were developed using different distance techniques (Random Forest or RF, Gradient Nearest Neighbour or GNN, and Most Similar Neighbour or MSN) and different combinations of response variables (model scenarios). Direct biomass imputation models were trained according to the biomass variables while indirect biomass imputation models were trained according to combinations of forest structure variables (e.g., basal area, stem density and stem volume of live and dead-standing trees). We also assessed the ability of our imputation method to spatially predict biomass variables across large areas in relation to a forest disturbance history over a 30-year period (1987–2016). Our results show that RF consistently outperformed MSN and GNN distance techniques across different model scenarios and biomass variables. The lowest error rates were achieved by RF-based models with generalized root mean squared difference (gRMSD, RMSE divided by the standard deviation of the observed values) ranging from 0.74 to 1.24. Whereas gRMSD associated with MSN-based and GNN-based models ranged from 0.92 to 1.36 and from 1.04 to 1.42, respectively. The indirect imputation method generally achieved better biomass predictions than the direct imputation method. In particular, the kNN model trained with the combination of basal area and stem density variables was the most robust for estimating forest biomass. This model reported a gRMSD of 0.89, 0.95 and 1.08 for total AGB, AGB of live trees and AGB of dead trees, respectively. In addition, spatial predictions of biomass showed relatively consistent trends with disturbance severity and time since disturbance across the time-series. As the kNN imputation method is increasingly being used by land managers and researchers to map forest biomass, this work helps those using these methods ensure their modelling and mapping practices are optimized.

Keywords: forest biomass; kNN imputation; Landsat time-series; forest disturbance

1. Introduction

Forest biomass is considered a key factor in carbon and water cycles in terrestrial ecosystems. Spatial estimates of forest biomass are therefore needed for understanding the sources and sinks of terrestrial carbon and for accomplishing scientific and practical tasks in forest management [1]. National and international reporting of forest biomass and carbon has traditionally relied upon field measurements from National Forest Inventory (NFI) programs [2]. However, these approaches cannot provide a continuous spatial distribution of forest biomass at the landscape scale, since field measurements often have limited spatial and temporal coverage, especially for large jurisdictions or remote areas [3]. To address this problem, researchers and practitioners often combine field measurements with remote sensing data to estimate forest biomass across large areas. Many studies have recently demonstrated the ability of lidar (light detection and ranging) data in providing high accuracy estimates of forest biomass and structure [4–10]. Lidar can be integrated with forest inventory data to make lidar-based biomass maps where data is available wall-to-wall [7–9,11,12], or to create lidar-based plots to support forest inventory where data are discrete available as sample transects [13,14]. Lidar-based plots can then be fused with larger coverage data, such as satellite imagery, to facilitate mapping forest biomass and structure at the land management scale [15–20]. However, while lidar is often available over forests in developed regions such as North America, such data is generally not available in developing regions, due to its high acquisition costs and advanced computational requirements. The immediate need of biomass estimations across large forest areas, therefore, relies on field-based inventories and multi-spectral remote sensing data provided by satellites such as Landsat and Sentinel.

The free availability of the entire historic Landsat archive (since 2008) makes it a popular remotely sensed data source for mapping current forest biomass, as well as monitoring forest biomass dynamics across space and time. The Landsat archive provides a 40-year collection of satellite imagery (since 1972) at a spatial resolution sufficient for capturing changes in forests [21]. This has facilitated the development of many approaches that utilize Landsat time-series imagery to characterize forest change [22,23]. Information from Landsat time-series has been widely used in many studies to estimate forest biomass and other structure attributes across large areas (e.g., [15,16,19,20,24–27]). At the conceptual level, Kennedy, Ohmann, Gregory, Roberts, Yang, Bell, Kane, Hughes, Cohen, Powell, Neeti, Larrue, Hooper, Kane, Miller, Perkins, Braaten and Seidl [24] developed a comprehensive forest biomass monitoring framework that is based on the analysis of Landsat time-series. Studies have also demonstrated the utility of Landsat time-series in improving forest biomass and structure estimates in comparison with methods that conventionally rely on single-date Landsat images [6,9,16,28]. For example, Pflugmacher, Cohen and E. Kennedy [9] indicated that the inclusion of spectral disturbance and recovery metrics extracted from Landsat time-series can improve biomass model results. More recently, Bolton, White, Wulder, Coops, Hermosilla and Yuan [16] demonstrated that time series metrics such as long-term Landsat spectral means and variability can also describe long-term forest dynamics. The inclusion of time series metrics not only improves the accuracy of empirical models but also makes spatial predictions of forest attributes more consistent with ecological changes such as those resulting from forest disturbance and recovery processes [19].

In forest mapping applications, k-Nearest Neighbour (kNN) imputation has been commonly used to leverage forest attributes of interest (response variables), derived from sample plot data, with spatial metrics (predictor variables), derived from remote sensing data, to generate spatial predictions of forest attributes (e.g., [15,16,19,24,29,30]). kNN imputation is a non-parametric and multivariate modelling method that aims to impute (or share) values of response variables from measured samples to target samples where response variables have not been observed (e.g., pixels covering an area of interest) [31]. Imputation associated with each target sample is based on the similarity (evaluated using a distance metric) between values of predictor variables associated with that target pixel and those associated with k nearest training samples. The imputed value of each target sample is assigned as the observed value of a particular training sample if $k = 1$, and as the average of observed values

associated to k nearest training samples if $k > 1$. Using k values greater than 1 often produces more accurate imputation results than using $k = 1$, but increases the bias/variance between imputed and observed values, which is an important consideration when comparing the performance of imputation models [31,32]. There are several common techniques to derive a nearest neighbour distance metric (hereafter, distance technique), including weighted Euclidean distance based techniques, such as Most Similar Neighbour (MSN) [33], or Gradient Nearest Neighbour (GNN) [34], and machine learning based methods such as Random Forest (RF) [35]. In fact, MSN is the configuration consisting of the canonical correlation analysis and $k = 1$ [33], and GNN is the configuration consisting of the canonical correspondence analysis and $k = 1$ [34]. However, several studies also used these techniques with $k > 1$ [36]. Some studies have compared kNN distance techniques [31,32,37]. Hudak, Crookston, Evans, Hall and Falkowski [31] compared several kNN distance techniques for imputing plot-level response variables (basal area and tree density) using airborne lidar data in small case study areas. The study found that the RF technique performed best in comparison with GNN and MSN. Eskelson, Temesgen, Lemay, Barrett, Crookston and Hudak [32] also found that RF outperformed other techniques when estimating current forest attributes from inventory data. As the kNN imputation method has been increasingly used in mapping forest biomass, it is critical for land managers and researchers to identify which distance technique performs better at regional and national scales.

Another important consideration when developing a kNN model is the method used for imputing biomass variables. This can be divided into two groups: direct and indirect imputation methods. In the direct method, forest biomass variables are included in kNN models as response variables and are thus directly imputed based on their relationship with predictor variables [30,37–40]. In the indirect method, however, kNN models are trained by response variables other than biomass [15,19,24]. These response variables are often forest structure attributes extracted from inventory plots, such as basal area, stem density and stem volume. Alternatively, variables may be measured from lidar-based plots such as mean vegetation height and percentage of returns [19]. Biomass variables are then attached as ancillary variables to the imputation predictions for target samples. That is, the nearest neighbours for each target sample are found based on the relationship between structure attributes and predictor variables and then are indirectly transferred to biomass variables. For instance, Kennedy, Ohmann, Gregory, Roberts, Yang, Bell, Kane, Hughes, Cohen, Powell, Neeti, Larrue, Hooper, Kane, Miller, Perkins, Braaten and Seidl [24], following Ohmann and Gregory [34], developed GNN imputation models to link a variety of plot-level measures including basal area by species and tree size classes with Landsat-based predictor variables. Using GNN models, each pixel was assigned tree measurements from an inventory plot, resulting in inventory-like maps, from which forest biomass was mapped. As the abovementioned studies were conducted in different locations and used different data sources, it is difficult to compare the imputation methods used for estimating forest biomass. Currently, there is no general agreement on whether direct or indirect imputation methods are better for estimating forest biomass from remote sensing data.

The main aim of this study is to determine the most accurate imputation approaches for mapping forest biomass at the landscape scale, using remote sensing and inventory data, and in doing so, fill current gaps in the literature. To achieve this, we compare the performance of several imputation approaches used for modelling aboveground forest biomass (AGB) over large areas using Landsat time-series and field plot measurements. In particular, we compare and contrast (1) direct and indirect biomass imputation methods and (2) three commonly used kNN distance techniques (RF, GNN and MSN), for predicting three forest biomass variables (total AGB, AGB of live-standing trees, AGB of dead-standing trees). In addition, we demonstrate and assess the utility of the kNN imputation method for spatially predicting biomass variables across large areas in relation to ecological changes (a 30-year history of forest disturbance, Nguyen, Jones, Soto-Berelov, Haywood and Hislop [23]).

2. Materials and Methods

The main steps implemented in this study are summarized in Figure 1. In summary, we derived a series of forest biomass (total AGB, AGB of live and dead-standing trees) and structure variables (e.g., basal area, stem density) from 633 forest inventory plots. As predictor variables, we calculated Landsat time-series based metrics across 19 Landsat tiles plus topographic and climatic variables. We extracted the predictor variables for each plot location and then developed, compared, and evaluated different kNN imputation approaches. Finally, we assessed the ability of the optimal kNN model to predict forest biomass in relation to disturbance history over a 30-year time-series (1987–2016). A more detailed description follows below.

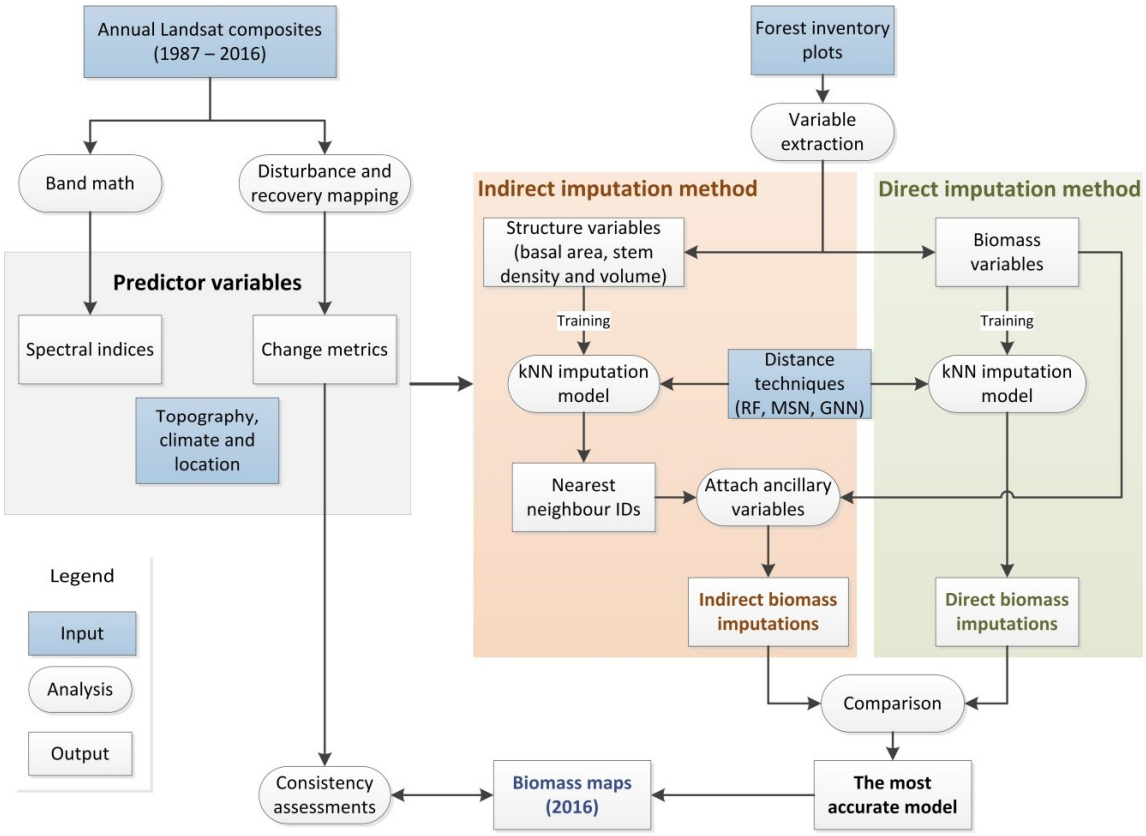

Figure 1. Overall flowchart of steps used for developing and comparing biomass imputation approaches.

2.1. Study Area

The study area contains the whole public land forest estate (7.1 million hectares) of Victoria, southeast Australia (Figure 2). Victorian public land forests extend across the state and include two main land tenures: state forests, and national parks and conservation reserves [41]. The area is stratified by eleven main bioregions (Interim Biogeographic Regionalization for Australia or IBRA, Figure 2), each of which has distinct ecological, geological and climatological features [41].

Victorian public forests include a wide range of ecosystems. North-western Victoria is mostly covered by mallee ecosystems, characterised by small woodland (up to 8 m tall) and multi-stemmed forests, on flat to undulating landscapes. In contrast, most ecosystems in the central part of the State are box-ironbark and red gum forests, which have dense to sparse canopies and reach up to 25 m in height. They are found on flat to undulating topography on rocky and auriferous soils. The most widespread and variable forest ecosystem in the study area consists of damp sclerophyll forests, which cover the central and eastern parts of Victoria. In this ecosystem, trees grow on loam, clay loam, and sandy loam soils and heights range from 40 m to 60 m. Distributed mostly in the eastern part of the state are wet

and dry sclerophyll forests. The former are the tallest forest ecosystems in Victoria, with trees reaching 75 m or more in height while the later include relatively low and spreading trees that reach a maximum height of 25 m [42]. Forests within the study area have been impacted by a series of disturbance events including fuel reduction burns, wildfires, logging and drought, resulting in significant changes in carbon biomass stocks [23].

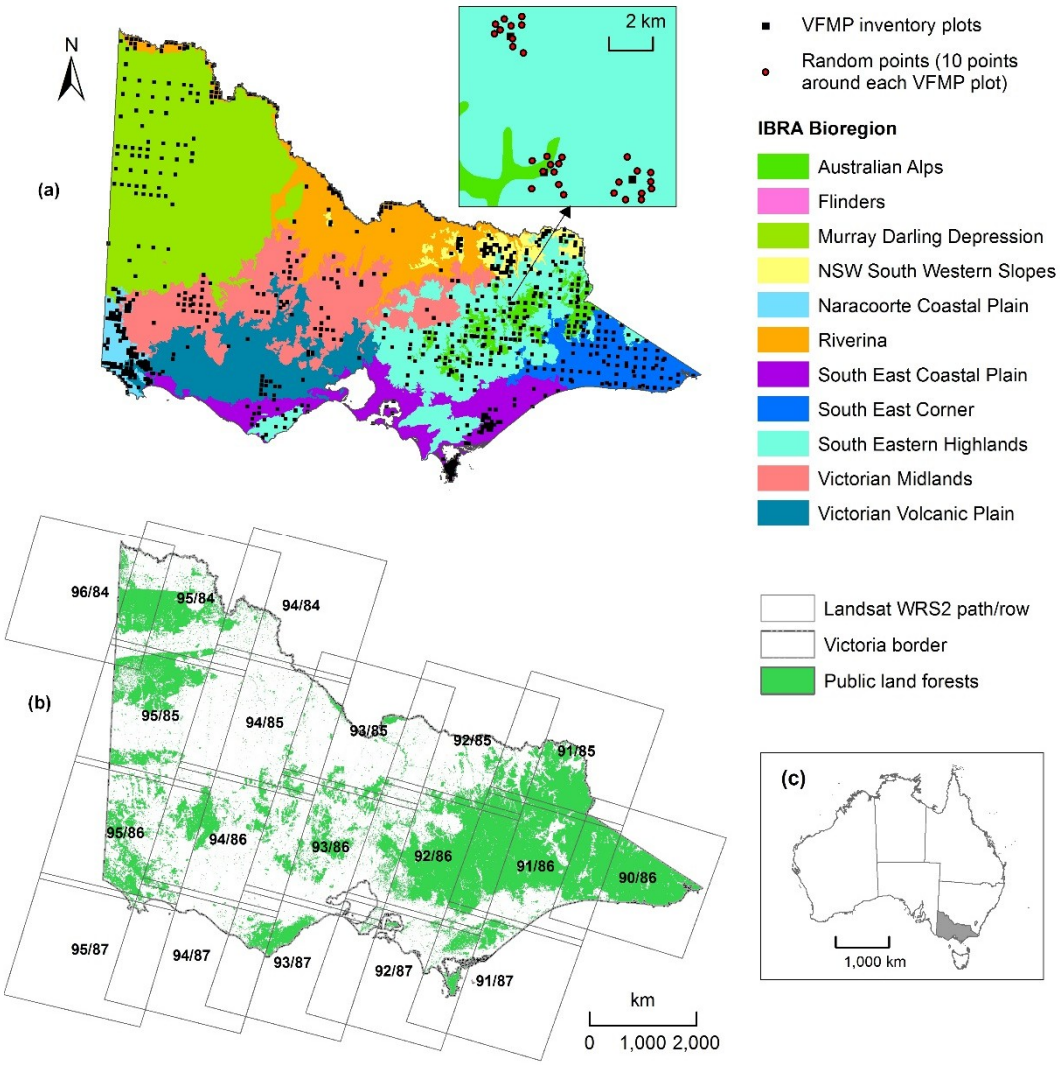

Figure 2. Study area in Victoria, Australia. (**a**) Bioregions and VFMP inventory plots with a local map showing examples of random points selected around an inventory plot; (**b**) public land forest extent and Landsat scenes; (**c**) map of Australia.

2.2. Forest Biomass and Structure Response Variables

Forest inventory plot data was collected as a part of the Victorian Forest Monitoring Program (VFMP) which is implemented by the Department of Environment, Land, Water, and Planning [43]. The VFMP includes a network of 786 permanent ground circular plots (0.04 ha) randomly distributed across a systematic statewide grid (Figure 2). In each plot, inventory data is collected on large trees, small trees, herbs, shrubs, and woody debris. Further descriptions of sample designs and field measurements can be found in Haywood, Mellor and Stone [43].

In this study, we extracted data from 633 VFMP plots measured between 2011 and 2016 and calculated nine tree-level forest biomass and structure variables (Table 1). Following Haywood and Stone [44], we estimated AGB (Mg·ha^{-1}) of all large live-standing trees (AGB_{live_tree}, ≥10 cm in

diameter at breast height—DBH) using a generic allometric equation for sclerophyll forests [45] of the form:

$$\ln(BLT_i) = -2.3267 + 2.4855 \times \ln(DBH_i), \quad (1)$$

where BLT_i is the AGB of large live-standing tree i. AGB of large dead-standing trees (AGB_{dead_tree}) was calculated by subtracting the amount of biomass in leaves, twigs and branches from AGB of live-standing trees. Total aboveground woody biomass (AGB_{total}) was calculated as the sum biomass of live- and dead-standing trees (AGB_{live_tree} and AGB_{dead_tree}), small trees (<10 cm DBH), stumps, slash and coarse woody debris (all fallen dead woody material, Haywood and Stone [44]). In addition to biomass variables, we calculated some structural measures that are summaries of tree measurements in each plot, including basal area (BA), stem density (TD, number of trees per hectare), and tree volume (VL) of both live- and dead-standing trees (Table 1).

Table 1. Forest biomass and structure variables extracted from inventory data.

Variable	Description	Mean (Range)	Unit
Biomass measurements			
AGB_{total}	Total aboveground woody biomass	284.9 (0.3–1037.7)	$Mg{\cdot}ha^{-1}$
AGB_{live_tree}	Total AGB of large live-standing trees	207.4 (0.1–907.8)	$Mg{\cdot}ha^{-1}$
AGB_{dead_tree}	Total AGB of large dead-standing trees	31.4 (0.0–349.8)	$Mg{\cdot}ha^{-1}$
Structure attributes			
BA_{live_tree}	Total basal area of live-standing trees	26.3 (0.3–140.9)	$m^2{\cdot}ha^{-1}$
BA_{dead_tree}	Total basal area of dead-standing trees	6.6 (0.0–134.9)	$m^2{\cdot}ha^{-1}$
TD_{live_tree}	Live-standing tree density	371.9 (25.0–2750.0)	Trees per hectare
TD_{dead_tree}	Dead-standing tree density	109.8 (0.0–2450.0)	Trees per hectare
VL_{live_tree}	Tree volume of live-standing trees	297.2 (0.5–2885.6)	$m^3{\cdot}ha^{-1}$
VL_{dead_tree}	Tree volume of dead-standing trees	20.0 (0.0–460.8)	$m^3{\cdot}ha^{-1}$

2.3. Predictor Variables

The candidate predictor variables included Landsat-based metrics as well as topographic and climatic ancillary data (Table 2). Landsat-based variables consisted of spectral indices and change metrics derived from the analysis of Landsat time-series.

Table 2. Predictor variables derived from Landsat time-series and topographic and climatic data.

Group	Variable	Description
Spectral indices	NBR	Normalised burn ratio
	TCB	Tasselled cap brightness
	TCG	Tasselled cap greenness
	TCW	Tasselled cap wetness
	TCA	Tasselled cap angle
	TCD	Tasselled cap distance
Change metrics	Pre-disturbance value	NBR value at the start vertex of disturbance segment
	Post-disturbance value	NBR value at the end vertex of disturbance segment
	Disturbance onset year	The year when disturbance begins
	Disturbance duration	Number of years between the start vertex and the end vertex of disturbance segment

Table 2. *Cont.*

Group	Variable	Description
	Disturbance magnitude	Difference in NBR value between the start vertex and the end vertex of disturbance segment
	Relative disturbance magnitude	Ratio of disturbance magnitude to pre-disturbance value
	Disturbance rate	Ratio of disturbance magnitude to disturbance duration
	Recovery onset year	The year when post-disturbance recovery starts
	Recovery duration	Number of years between the start vertex and the end vertex of recovery segment
	Recovery magnitude	Difference in NBR value between the start vertex and the end vertex of recovery segment
	Relative recovery magnitude	Ratio of recovery magnitude to post-disturbance value
	Recovery rate	Ratio of recovery magnitude to recovery duration
	Time since disturbance	Number of years since disturbance ends
	Disturbance level	High, medium, or low disturbance
	Disturbance causal agent	Fire, logging, and other (drought, insects, flood)
Topographic and climatic metrics	Elevation	Elevation in meters
	Slope	Slope in degrees
	Precipitation	Mean total rainfall
	Temperature	Mean annual temperature
Location	X	Northing
	Y	Easting

2.3.1. Landsat Time-Series

The study area is covered by 19 Landsat WRS-2 tiles ranging from row 84 to 87, and path 90 to 96 (Figure 2). From the USGS archive, we processed all available Level-1 Terrain Corrected (L1T) Landsat TM/ETM+ images acquired within a pre-defined southern hemisphere summer window (from December to the end of February) from 1987 to 2016. Annual anniversary-date, best observation mosaic composites were created for the 30-year time period using all cloud-free observations within the summer window. For more details on image processing, please see Nguyen, Jones, Soto-Berelov, Haywood and Hislop [23]. For each composite of surface reflectance, we calculated the Normalized Burn Ratio index (NBR) [46], and the Tasselled-cap (TC) components of Greenness (TCG), Brightness (TCB), and Wetness (TCW) [47]. We also computed TC angle (TCA = arctan (TCG/TCB), Powell, Cohen, Healey, Kennedy, Moisen, Pierce and Ohmann [38]) and TC distance (TCD = $\sqrt{\text{TCG}^2 + \text{TCB}^2}$, Duane, et al. [48]).

To derive spectral trends for each pixel, we analysed annual NBR time-series using the LandTrendr temporal segmentation algorithm developed by Kennedy, et al. [49]. Similar to other studies (e.g., [49–52]), we found that NBR was the most sensitive spectral index for capturing forest disturbance from Landsat time-series [53]. The core segmentation process includes two steps: finding vertices and fitting trends. The first step establishes the vertex years that define temporal breakpoints, reducing

year to year noise over the time series. The second step determines the best straight-line trajectory that fits through those vertices, resulting in a fitted spectral trajectory for each pixel within the processing area (Figure 3).

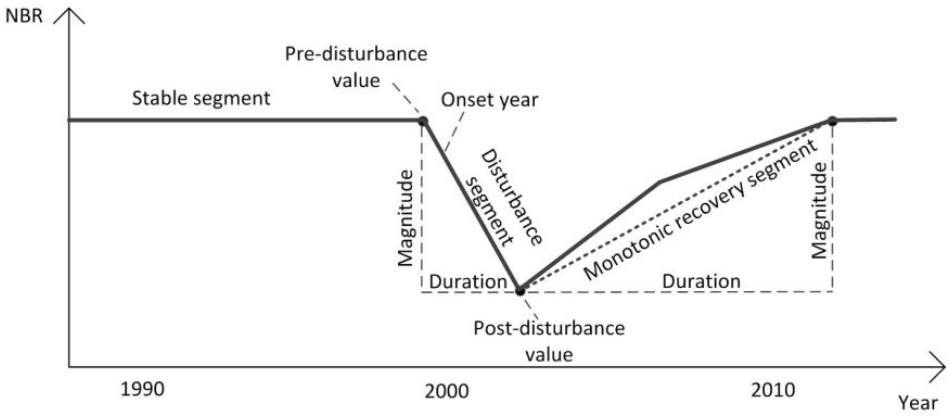

Figure 3. Example of a trajectory of NBR time-series and change metrics.

From the NBR fitted trajectories, we computed a suite of change metrics representing disturbance and recovery trends for each pixel (Table 2). We first labelled the greatest disturbance segment which corresponds with the greatest negative change magnitude. We then identified the subsequent recovery segment as the monotonic positive segment following the greatest disturbance (Figure 3). More details can be found in Nguyen, Jones, Soto-Berelov, Haywood and Hislop [23]. These two segments were then used to derive disturbance and recovery metrics that represent year, duration, and spectral magnitude of change, as well as pre- and post-change spectral conditions [19,20,23,50]. We also calculated the number of years since the greatest disturbance (or time since disturbance, TSD) to the last year of the time-series (2016). However, for this metric, we applied an additional rule that set the TSD of all un-changed pixels to 50 years. TSD is an important predictor of the regrowth of forests following disturbance. However, since we only had disturbance data for the last 30 years, we assigned non-disturbed pixels a value of 50, to represent mature forests. This was done so as not to confuse the model with a value of 0, for example, which might be interpreted incorrectly. Although 50 was a somewhat arbitrary value, in these forests, it was considered sufficient to represent mature forests. Using disturbance and recovery metrics, we developed classification models to predict disturbance severity levels (high, medium and low; with a high level disturbance resulting in the full removal of trees in forests) and associated causal agents [23]. These data were also included as predictor variables in biomass models in this study, resulting in 15 change metrics in total. A spatial filtering process was applied to the derived change metrics to select only pixels within forested areas and remove change events smaller than 0.5 ha [23].

2.3.2. Topographic and Climatic Ancillary Data

Topographic variables including elevation and slope were derived from the Shuttle Radar Topography Mission (SRTM) 1 arc-second resolution (~30 m) dataset [54]. Mean annual temperature and mean total rainfall were extracted from the Global Climate Data (WorldClim) with a spatial resolution of 1 km [55]. These datasets were obtained and resampled to a spatial resolution of 30 m, to match that of Landsat.

2.4. Biomass Model Development

2.4.1. Variable Extraction

For each inventory plot location, we extracted the values from the prepared predictor variables (Table 2) associated with the single Landsat pixel that contained the plot centre. Although extracting

values from a kernel size, such as a 3 × 3 pixel window, minimises the spatial mismatch between spatial data and inventory plots, the use of the single Landsat pixel allows us to capture the specific disturbance history associated with each inventory plot. Earlier research has shown that a disturbance may impact pixels within a plot in different ways [23]. For example, a fire may have a higher severity in a given pixel than its adjacent pixels. To reduce the mismatch between plot condition at the time of field measurement and time of image acquisition, the Landsat-index variables were extracted from the composite images with the dates that closely coincided with plot measurement dates. Values of change metrics for each disturbed plot were calculated for the time period from 1987 to the plot observation year. In addition, we identified and removed potential outliers to improve data quality. Outliers were defined as plots associated with edge effects such as adjacent water or roads. This removed approximately 8% of the of inventory plots.

2.4.2. Variable Selection

As reducing the number of predictor variables can avoid detrimental impacts on kNN imputation accuracy, the preliminary modelling step was to determine an optimal set of predictor variables. To achieve this, we first ran the least absolute shrinkage and selection operator (LASSO, Efron, et al. [56]) model to rank all predictor variables based on their importance to each response variable. The LASSO model quantifies the strength of the relationship between predictor variables and response variables using a pseudo R^2 metric that ranges from 0 to 1. The importance rankings of predictor variables for individual response variables were made based on this metric. Predictor variables with consistently low rankings were excluded from further analyses. Redundant variables (or highly correlated variables) were also identified and removed by calculating Pearson correlation coefficients between all pairs of remaining variables and removing those with $r > 0.9$.

2.4.3. Imputation Models

We developed and compared different kNN (with $k = 1$) imputation approaches to predict three biomass measurements (AGB_{total}, AGB_{live_tree}, AGB_{dead_tree}) based on several combinations of response variables and distance techniques. For each of the three distance techniques (RF, GNN, or MSN), we developed six model scenarios using different groups of response variables (Table 3), resulting in 18 kNN models in total. The first model scenario (BM) was the direct biomass imputation model since it was trained by biomass response variables (AGB_{total}, AGB_{live_tree} and AGB_{dead_tree}). The nearest neighbour was found by directly relating observed biomass variables to predictor variables. In contrast, the other five model scenarios (BA, TD, VL, BA-TD and VL-TD) were the indirect biomass imputation models, as the nearest neighbour was found based on the relationships between predictor variables and forest structure variables rather than biomass variables. Biomass measurements of the corresponding training plots were not included in these models but were subsequently attached as ancillary variables to impute each target pixel. The combination of BA and VL was not included in our analysis since these variables are highly correlated. The Pearson correlation coefficient between BA_{live_tree} and VL_{live_tree} was 0.91 and between BA_{dead_tree} and VL_{dead_tree} was 0.96.

GNN and MSN identify the nearest neighbour based on weighted Euclidean distance techniques. The MSN technique computes the distance in projected canonical space while the GNN technique computes distance using a projected ordination of predictors based on canonical correspondence analysis (CCA) [33,34]. The distance metric in RF, on the other hand, is derived based on a proximity matrix [35]. The elements of the proximity matrix contain the proportion of decision trees where both training and target samples are found in the same terminal node. The statistical distance metric is calculated as one minus that proportion [57]. After testing with different model parameters, we set the number of trees (*ntree*) to 200 for each RF model (associated with each response variable), and the number of predictor variables selected at each node (*mtry*) to the default, based on the square root of the number of predictors. These values were chosen since they minimized the model errors (RMSE) within the training dataset. Since kNN imputation creates multiple RF models associated with the

number of response variables and shares the number of trees across the individual models, the input parameter of *ntree* of kNN model was actually 200 multiplied by the number of response variables. We built our models using the R-package *yaImpute* [57].

Table 3. Model scenarios developed for each distance technique (BM = biomass, BA = basal area, TD = stem density, VL = tree volume, X = denotes response variable group in each model).

Response Variables	Model Scenarios					
	BM	**BA**	**TD**	**VL**	**BA-TD**	**VL-TD**
Biomass variables						
AGB_{total}	X					
AGB_{live_tree}	X					
AGB_{dead_tree}	X					
Structure variables						
BA_{live_tree}		X			X	
BA_{dead_tree}		X			X	
TD_{live_tree}			X			X
TD_{dead_tree}			X			X
VL_{live_tree}				X	X	X
VL_{dead_tree}				X	X	X

2.4.4. Model Evaluation

We evaluated the accuracy of each model scenario using a leave-one-out cross validation approach. For each sample plot, models were trained by all data except the candidate plot which was then treated as a target observation. Errors were computed for each withheld sample and averaged to evaluate model performance. For each biomass variable, imputed values were compared to observed values using the generalised root mean square difference (gRMSD, RMSE divided by the standard deviation of the observed values under the assumption that they are representative of the population, Crookston and Finley [57]), and relative mean deviation (rMD, Gorard [58]) which is a measure of bias:

$$\text{rMD} = \frac{\frac{1}{n}\sum_{i=1}^{n}|\widetilde{x}_i - x_i|}{|\overline{x}|}, \quad (2)$$

where $\widetilde{x}_i$ and x_i are the imputed (predicted) and observed biomass value, respectively, of the ith sample, and $\overline{x}$ is the mean of observed values.

2.4.5. Assessment of Biomass Imputation Maps Using Disturbance History

After evaluating the accuracy metrics, we selected the kNN model that consistently performed well across the biomass response variables, to implement spatial imputations (k = 1) to forested pixels across the study area for the year 2016. We assessed the ability of the selected kNN model to predict forest biomass in relation to disturbance history throughout the 30-year time-series (1987–2016). To facilitate this, we created a reference dataset containing 7860 reference points (with a minimum distance of 250 m between them) across the study area [59]. These points were built around the VFMP inventory plot network (10 points around each plot), thus they were also stratified according to bioregions (Figure 2). Points that fell on the boundary of land cover types, or at the edge of disturbances were shifted away from the edge to avoid mis-registration errors. Reference points were then interpreted and attributed with disturbance severity levels (high, medium and low), and associated causal agents (fire, logging and other) using a multiple-lines of evidence approach. For a detailed explanation of the multiple lines of evidence approach, see Soto-Berelov, Haywood, Jones, Hislop and Nguyen [59]. Some points that fell in non-forest areas, defined following Nguyen, et al. [60], were removed from the dataset.

From the kNN imputed maps, we extracted the mean values from a 3 × 3 pixel window around each reference pixel, for each of the biomass response variables. As the range of biomass often varies across forest ecosystems, it is not appropriate to compare reference points from different bioregions using imputed biomass values. Thus, we grouped the random points by bioregion and then scaled biomass values for each set from 0 to 100 (corresponding with low to high biomass in each bioregion).

To examine the relationship between scaled biomass values and recent disturbance events, we selected reference points that had been disturbed by fire and logging events that occurred between 2013 and 2016. The main reason for choosing this time period is that, for non-stand replacing disturbance events occurring in over three years prior to the mapping date (2016), the evidence of their impacts on biomass maps may not be clear as forests may have regrown [61]. Trends and variation of biomass associated with these points were then analysed according to disturbance severity level and causal agent using boxplots. In addition, we assessed biomass patterns according to time since disturbance or TSD, for each disturbance severity level. This analysis was completed for all disturbances occurring within the time-series (1987–2016). We first stratified TSD into nine intervals (0–3, 3–6, 6–9, 9–12, 12–15, 15–18, and 18–26 years). The basis of this division was to illuminate the trend of biomass according to TSD. A shorter interval (3 years) was applied for TSD smaller than 18 years since forests often significantly change/grow during this period. We then grouped the disturbance points using these TSD intervals and by disturbance severity levels, and calculated the mean of the scaled biomass values for each group.

3. Results

3.1. Variable Selection

The LASSO model reported that spectral indices (excepting TCB and TCD), climatic and topographic variables were the most important predictor variables related to our response variables (Figure 4). Pre- and post-disturbance values and relative disturbance magnitude were the most important change metrics. Variables such as disturbance and recovery onset year and duration had consistently low importance rankings and thus were excluded from kNN biomass models. Although disturbance severity levels and causal agents had relatively low importance, these variables were included in biomass models since they have been effective in other studies [15,16]. TCB and TCD were relatively important but were excluded as redundant variables, based on having high correlations. Nineteen remaining variables were finally selected to use in the kNN biomass models.

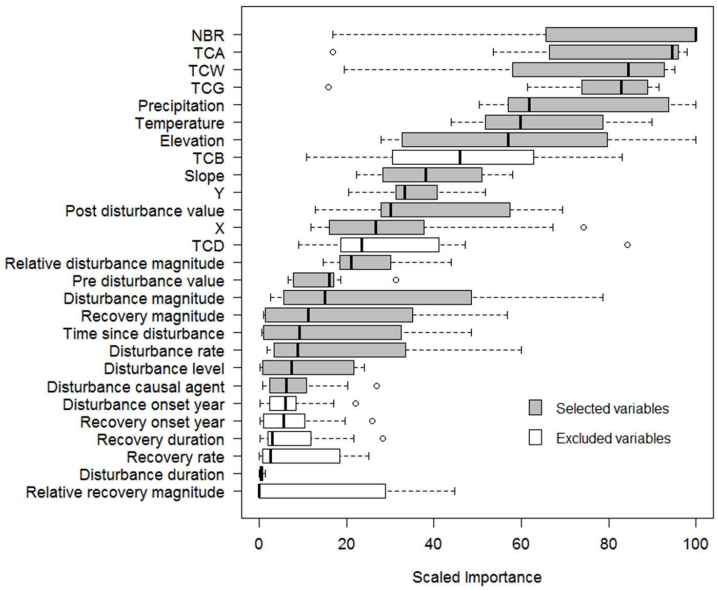

Figure 4. Scaled importance of predictor variables to response variables, reported by the LASSO model.

3.2. Biomass Imputation Model Accuracy

The results of biomass model assessment varied across model scenarios, imputation techniques (RF, GNN and MSN) and the biomass variables (AGB_{total}, AGB_{live_tree} and AGB_{dead_tree}). Figure 5 shows the pattern of gRMSD across all tested models. It clearly indicates that RF consistently outperformed MSN and GNN distance techniques in terms of gRMSD. Across biomass models, AGB_{total} consistently achieved lower gRMSD values (ranging from 0.74 to 1.34) than AGB_{live_tree} (0.88–1.39) and AGB_{dead_tree} (1.08–1.42). The lowest gRMSD values across biomass variables were reported by RF-based BM, BA and VL models, while the highest values were reported by GNN-based BA and TD models. In addition, model scenarios performed differently across imputation techniques. For example, the BM model scenario achieved the lowest errors when using RF and MSN but not when using GNN. The BA model scenario showed better performance than the other model scenarios when using RF but it was one of the worst performings when using GNN. The VL-TD model scenario reported lower error rates than the other model scenarios when using MSN or GNN techniques but higher when using RF. TD was the worst performing model scenario regardless of distance techniques.

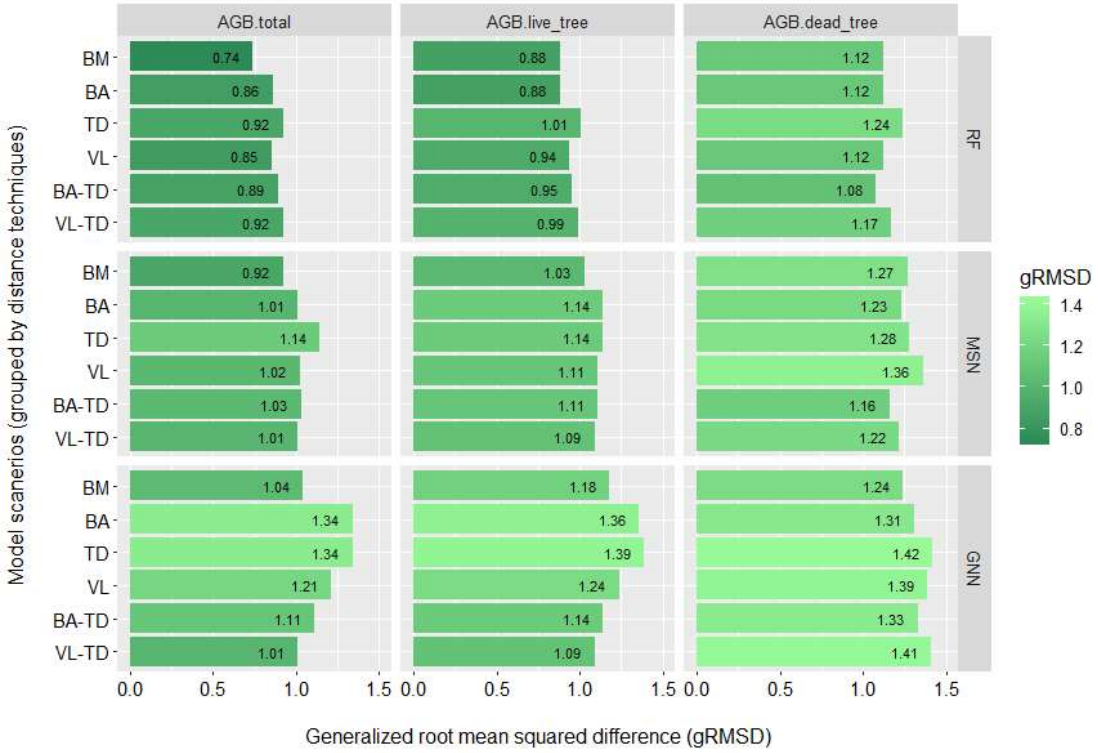

Figure 5. Generalized root mean squared difference of biomass imputations reported by kNN models (BM = biomass, BA = basal area, TD = stem density, VL = tree volume).

The patterns in model-strength indicated by rMD were relatively similar to those illustrated by gRMSD (Figure 6). RF performed much better than the other distance techniques and rMD values were generally lower for AGB_{total} and AGB_{live_tree} (ranging from 0.26 to 6.44 for AGB_{total}, from 0.03 to 6.46 for AGB_{live_tree}, and from 0.55 to 6.45 for AGB_{dead_tree}). The lowest values associated with AGB_{total} and AGB_{live_tree} were achieved by the RF-based BA model while the lowest value associated with AGB_{dead_tree} was obtained by the RF-based TD model. The highest values were reported by GNN-based BA and TD models. Similar to gRMSD, the patterns of rMD indicate that model scenarios show varied performances across distance techniques and biomass variables. When using RF, the BA model scenario reported lower error rates for AGB_{total} and AGB_{live_tree} (0.26 and 0.03, respectively) than the other model scenarios. When using MSN and GNN distance techniques, however, lower error rates were reported by TD and VL-TD model scenarios, respectively. In addition, models associated with

low bias for AGB_{total} and AGB_{live_tree} often reported relatively high bias for AGB_{dead_tree}, for example 4.84 for the RF-based BA model and 4.42 for the MSN-based TD model.

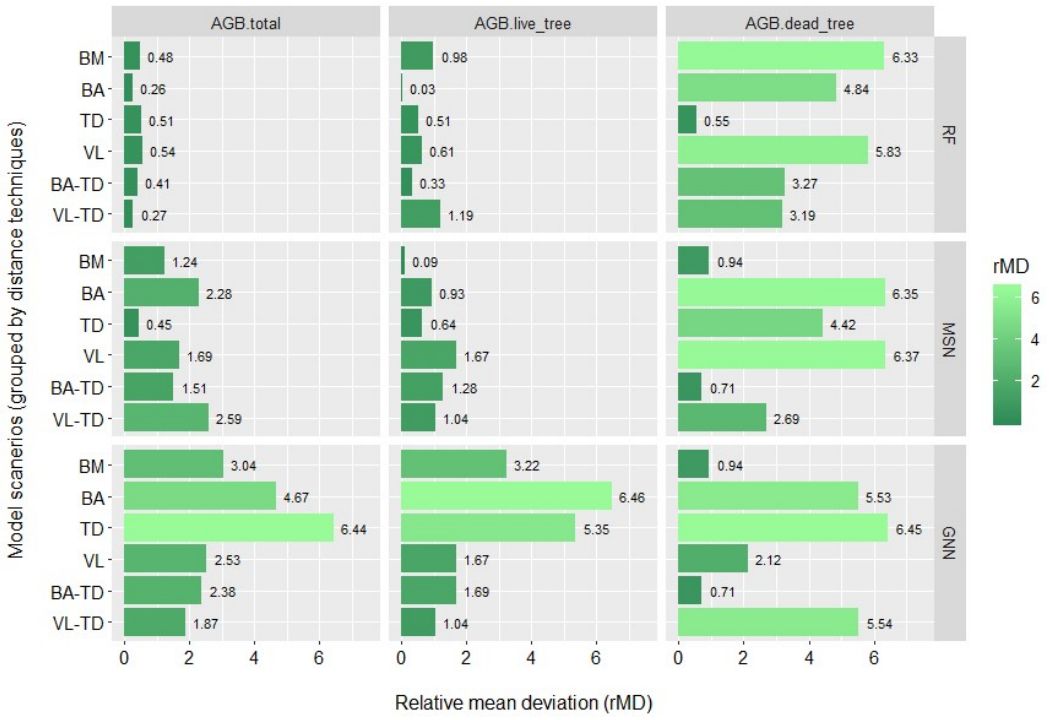

Figure 6. Relative mean deviation of biomass imputations reported by kNN models (BM = biomass, BA = basal area, TD = stem density, VL = tree volume).

3.3. Biomass Imputation Maps in Relation to Disturbance History

Based on model comparisons, we selected the model trained by the combination of basal area and stem density variables (RF-based BA-TD model) to produce imputation maps of biomass for 2016, across the study area (Figure 7). This model reported relatively and consistently low error rates across both accuracy metrics (gRMSD and rMD) and all three biomass variables (Figure 8). For all biomass variables, the model over- and under-predicted low and high observed biomass values, respectively. The imputation maps of forest biomass show clear longitudinal trends at the state level, with lower biomass in northwestern mallee forests and higher in southeast sclerophyll forests (Figure 7a). At the local scale, biomass predictions are spatially consistent with ground conditions of forests, accurately capturing evidence of recent forest disturbance (Figure 7b–e).

Predictions of biomass showed relatively consistent trends in relation to the severity of recent (2013–2016) fire and logging disturbance events (Figure 9). For all three biomass variables, high severity disturbance consistently resulted in low biomass predictions. Scaled biomass values were generally higher when relating to lower disturbance levels (medium and low severity), excepting small trends in AGB_{total} and AGB_{live_tree} associated with low severity logging. In addition, fire disturbed areas indicated slightly lower scaled biomass values in comparison with logged areas.

The trends of biomass predictions in relation to TSD varied across disturbance severity levels and biomass variables (Figure 10). Similar to the analysis on recent disturbances, the analysis on all disturbances within the 30-year time-series indicated that biomass predictions associated with low severity disturbance were generally higher than those associated with medium and high severity disturbance. For high and medium severity levels, scaled biomass values were generally higher with increased TSD. For the low severity level, on the other hand, scaled biomass values were generally stable across TSD intervals. However, these trends were not linear. There was a substantial increase in biomass predictions within the 3–6 and 6–9 years TSD intervals across all disturbance severity levels and biomass variables.

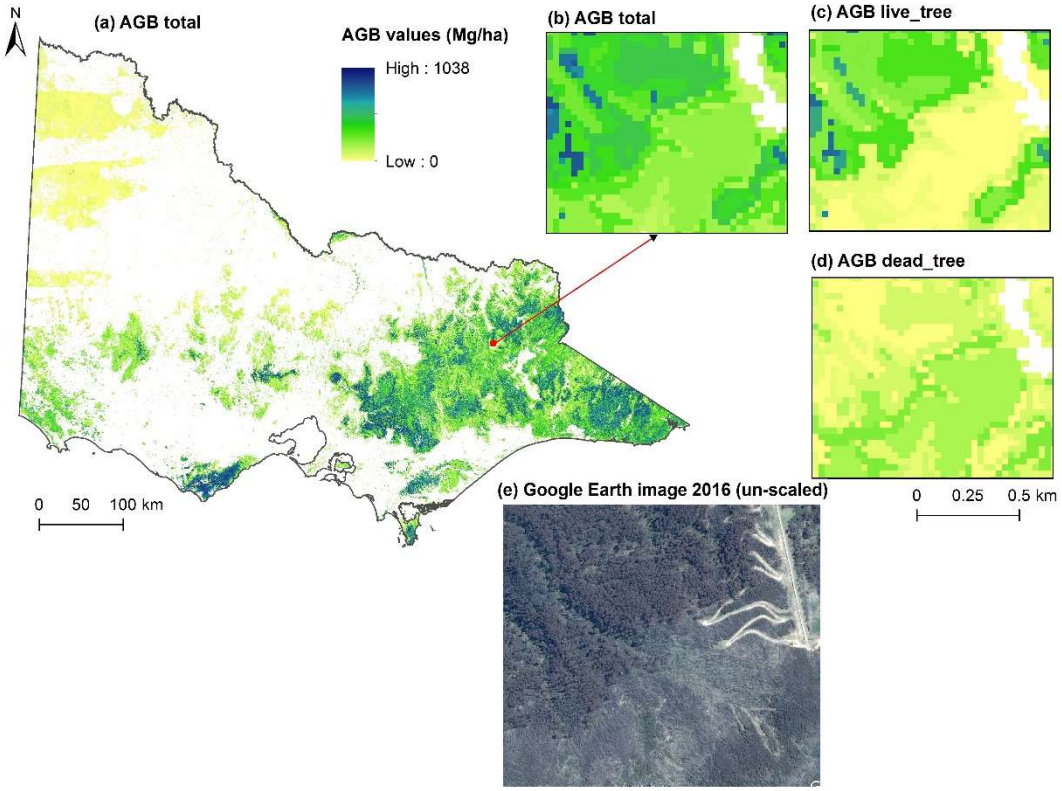

Figure 7. (**a**) The imputation map of total AGB for 2016 across public land forests in Victoria. (**b**–**d**) Imputation maps of total AGB, AGB of live tree and dead tree, respectively, at a local scale. (**e**) A Google Earth image (un-scaled) showing the same area as the local maps. Predictions of AGB are consistent with forest conditions displayed on the Google Earth image, with total AGB at a medium level. Live trees predominate in the top-left corner while dead trees are dominant in the bottom-right corner as a consequence of a 2007 fire.

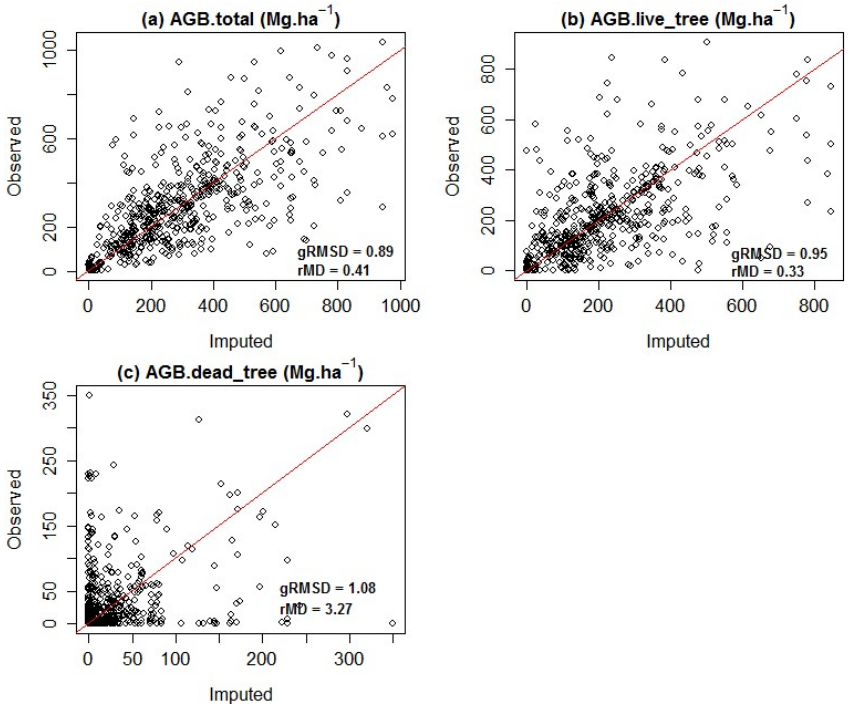

Figure 8. Imputed versus observed biomass values from leave-one-out cross validation, reported by the RF-based BA-TD model.

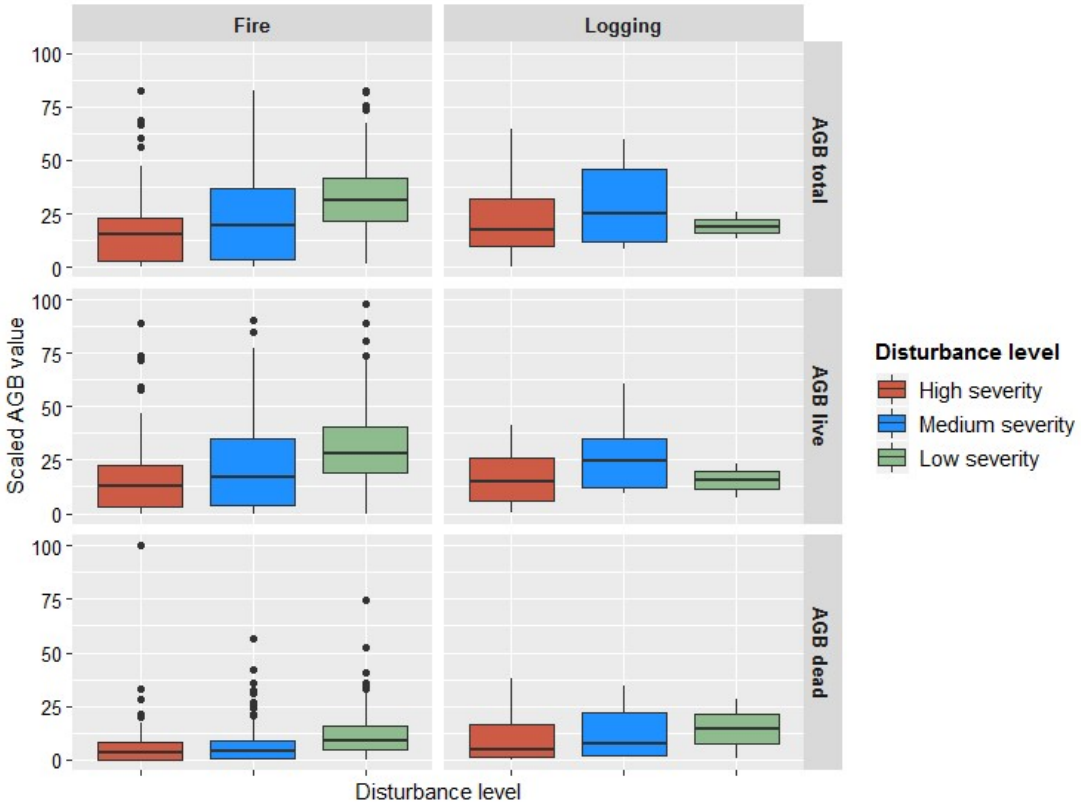

Figure 9. Boxplots of scaled imputed AGB values by different disturbance severity levels associated with fire and logging disturbances occurring between 2013 and 2016.

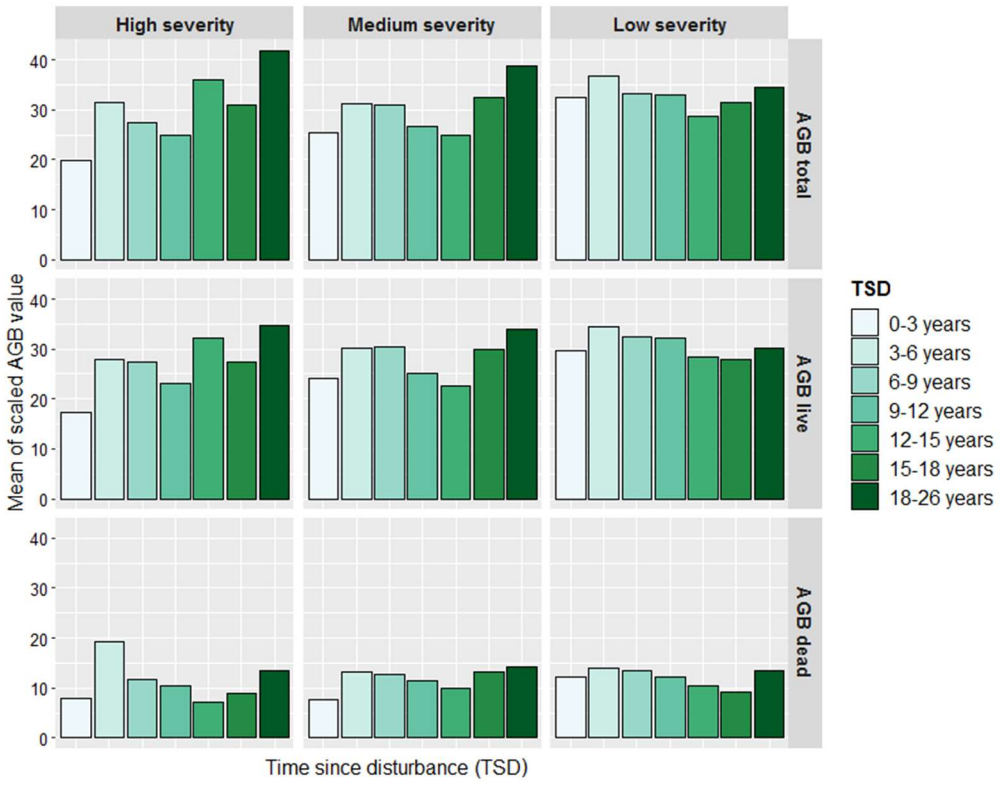

Figure 10. Trends of scaled imputed AGB according to disturbance severity and time since disturbance (TSD).

4. Discussion

Studies have demonstrated different kNN imputation approaches for the empirical estimation of forest AGB using remotely sensed time-series and forest inventory data. However, different approaches can produce markedly different results. Thus, knowledge of which approach is the most appropriate is needed for land managers and researchers. In this study, we compared several kNN-based imputation approaches for estimating forest biomass using Landsat time-series and field plot measurements. Specifically, we evaluated three commonly used kNN distance techniques: RF, MSN and GNN, and two biomass imputation methods: direct and indirect imputation. While a few studies have compared the performance of different distance techniques used in kNN imputation models [31,32], none of them conducted their comparisons by leveraging multispectral time-series data and forest inventory data. Furthermore, there is no existing literature demonstrating whether biomass variables should be included as response variables and directly imputed, or be indirectly imputed from models built upon other structure variables. In addition, we evaluated the utility of kNN imputation models built upon Landsat time-series and inventory data by conducting ecological validations of imputed biomass maps.

Our results indicate that the accuracy of kNN biomass imputation models (k = 1) varies with different distance metrics. Models based on the RF distance technique, which calculates the distance metric based on a proximity matrix, generally outperformed those based on MSN and GNN distance techniques. As shown in Figures 5 and 6, RF-based models consistently achieved lower error rates of biomass imputation (in both gRMSD and rMD values) as compared to MSN and GNN-based models. These results agree with previous studies that compared imputation techniques for modelling forest attributes [31,32]. In addition, we found the performance of RF-based models to be relatively stable while that of MSN-and GNN-based models tended to be inconsistent across varied model scenarios. The rMD value associated with AGB_{total} reported by RF-based models ranged from 0.26 to 0.54, while the range reported by MSN and GNN-based models were significantly higher (from 0.45 to 2.59 and from 1.87 to 6.44, respectively, Figure 6).

The results also indicate the influence of the number of response variables included in kNN biomass imputation models on the performance of different distance techniques. In general, RF-based models were not significantly impacted by the number of response variables, given the models with four response variables (RF-based BA-TD and VL-TD models) achieved comparable accuracies to those with two variables (such as RF-based BA and VL models). This could be due to the small number of response variables (2 to 4) included in our RF-based models. Previous studies used a larger number of variables and found that RF works optimally with few variables and when factors are used rather than continuous values [31,57]. In contrast to RF, GNN-based models performed significantly better with an increased number of response variables. Error rates reported by GNN-based models were highest for models with two response variables (such as GNN-based BA and TD models) and lowest for those with four response variables (GNN-based BA-TD and VL-TD models). This compares favourably with other studies [34,62]. The performance of MSN-based models varied with the number of response variables, making it difficult to identify a specific trend. As our models included a relatively small number of response variables, further work is needed to investigate the impact of the number of response variables on the performance of kNN imputation with different distance techniques.

Among indirect biomass imputation model scenarios (BA, VL, TD, BA-TD, and VL-TD), models trained by basal area or stem volume variables generally achieved better accuracy than those trained by stem density variables (Figures 5 and 6). This was expected, since basal area and stem volume variables are often more correlated with the biomass variables than stem density variables. However, models with only basal area or stem volume response variables (BA and VL scenarios) often produced unbalanced accuracies across biomass ranges, exhibiting high bias for the dead biomass variable (AGB_{dead_tree}, Figure 6). The inclusion of stem density variables (BA-TD and VL-TD scenarios) significantly reduced this bias, balancing accuracy across all biomass variables. rMD values associated with AGB_{dead_tree} reported by BA and VL models ranged from 2.12 to 6.35 while those reported by

BA-TD and VL-TD models ranged from 0.71 to 3.27 (with an exception of 5.54 from GNN-based VL-TD model).

The determination of whether direct or indirect imputation method is better for forest biomass estimation depends on the distance technique used in the kNN model. Although the direct biomass imputation model (BM model scenario) performs relatively well across different distance techniques, it is not always the best method for estimating biomass variables. When the MSN distance technique is applied, the BM model scenario performed better overall than the indirect model scenarios given it consistently reports lower errors for all three biomass variables. The next best is the BA-TD model scenario which produced slightly higher errors for total and live tree AGB but lower errors for dead tree AGB (Figures 5 and 6). When using the RF distance technique, however, the BM model scenario did not perform as well as the BA and BA-TD scenarios. Although the BM model scenario achieved the lowest gRMSD values (0.74 and 0.88 for AGB_{total} and AGB_{live_tree}, respectively), it produced greater bias than the BA and BA-TD model scenarios. In particular, the rMD values reported by the BM model scenario ranged from 0.48 to 6.33 while those reported by BA and BA-TD were from 0.03 to 4.84 and from 0.33 to 3.27, respectively (Figure 6). Despite a relatively high rMD value for AGB_{dead_tree} (4.84), it is reasonable to consider BA as the best model scenario for imputing AGB variables when applying the RF distance technique. The results from GNN-based models indicate the superior performance of VL-TD and BA-TD model scenarios. The model trained by stem volume and stem density variables (VL-TD) obtained the lowest errors for total and live tree AGB but high errors for dead tree AGB. Whereas, the model trained by basal area and stem density variables (BA-TD) achieved more consistent results across all three biomass variables (Figures 5 and 6). These results suggest that BA-TD is the most robust model scenario for imputing forest biomass given it maintains the most consistent performance across the tested distance techniques and biomass variables. Overall, the indirect imputation method, particularly kNN models trained using a combination of basal area and stem density variables, achieved better biomass estimates than the direct imputation method.

It is important to note that we developed and compared the kNN models using the single nearest neighbour (k = 1). This makes the comparison consistent and our methods and results applicable in other study areas/contexts. While increasing k (to an optimal value) reduces the imputation error, the determination of an optimal k value is often difficult and depends on many factors including distance metrics, response variables and forest environments [32,36]. Maintaining consistent parameters (i.e., k value) and methods allows the imputation results to be evaluated more effectively [32]. We note also, that the use of a single nearest neighbour has been increasing in forest applications with kNN models, particularly in biomass estimation [15,16,19,29,36,63]. Chirici, Mura, McInerney, Py, Tomppo, Waser, Travaglini and McRoberts [36] found in their review work that k = 1 is the most common selection to use with MSN and GNN techniques. This is reasonable since these techniques are initially created to use with the single nearest neighbour, as mentioned in the introduction. Recent studies using the RF technique also often selected k = 1 for their imputation models [16,19,63]. As forest inventory programs are increasingly developed systematically, measurements from inventory plots are representatives of forest populations [64]. The use of k = 1 is thus recommended to keep variance in the imputations similar to variance in the observations. Although further work is required to examine how higher k values impact the comparison results, we strongly believe that RF would outperform MSN and GNN distance techniques, regardless of the k values used.

Our results indicate that RF was the most accurate distance technique, thus the selection of the best kNN imputation model for mapping biomass variables was between the RF-based BA and BA-TD models. The former was most accurate for AGB_{total} and AGB_{live_tree} estimates, while the latter resulted in a more balanced performance across all three biomass variables. Our biomass maps were predicted using the BA-TD model as we aimed to focus on both live and dead biomass pools (Figure 8). Results from the model showed that AGB_{total} and AGB_{live_tree} achieved better accuracies than AGB_{dead_tree}. This supports the results of other studies, which have demonstrated that the total

and live biomass variables often have better relationships with Landsat spectral values than dead biomass [6,9,65]. Spatial predictions of biomass were consistent with the distribution of different forest systems across the state (as described in Section 2.1), with low productivity in the low-spare mallee forests in the northwest and high productivity in the high-dense sclerophyll forests in the southeast (Figure 7).

Predictions of biomass were relatively consistent with the 30-year disturbance history of forests within the study area (Figures 9 and 10). Forest dynamics within the study area are dominated by fire and logging disturbances. When relating current predictions of forest biomass with recent disturbance events, we found that increased disturbance severity was consistently associated with decreased biomass predictions (Figure 9). The reduction of dead tree AGB after a high severe fire is expected given we defined a high level disturbance as the full removal of trees in forests [23]. This trend was also evident in the analysis based on all disturbances occurring from 1988 to 2016 (Figure 10) and is consistent with findings from other studies [19,26]. In addition, the results also suggest that predicted biomass was more sensitive to fire than to logging disturbance (Figure 9). This should be expected as un-wanted parts of logged trees (such as branches and stumps) and small trees often remained in forests after a selective logging event. Although current predictions of biomass are more variable when relating to time since disturbance (TSD), the trends were relatively consistent with post-disturbance forest recovery. Biomass values were often lowest within 0 to 3 years following a disturbance and generally reached an asymptotic level at 18–26 years after disturbance (Figure 10). Within Victorian forests, fires often cause high rates of tree mortality, which can exist for many years after a fire, resulting in increased trends in dead tree biomass [44]. In general, biomass predictions showed relatively high values within 3–9 years after a disturbance. This trend was consistent across the biomass variables and disturbance severity levels (Figure 10). The reason for this could be that Landsat spectra and indices are known to saturate at relatively low leaf area and biomass levels. These can be attained only a few years post-disturbance [61]. This also suggests that the uncertainty of predicted biomass maps can be informed by forest disturbance history.

Our analysis on variable selection further clarifies the benefits of including change metrics from Landsat time-series to improve predictions of forest biomass. To our knowledge, this is the first time spatial change metrics extracted from Landsat data have been combined with the systematic network of forest inventory, across large areas of sclerophyll forests in Victoria, Australia. Our results from the variable importance analysis were consistent with those from previous studies [6,9,16,19,20,63]. Spectral indices such as NBR and TCA were the most important variables overall, and change metrics such as disturbance and recovery magnitude, and TSD, were particularly important for modelling dead biomass and structure variables. Similar to Zald, Wulder, White, Hilker, Hermosilla, Hobart and Coops [19], our results indicate that change attribution variables (disturbance level and causal agent) are less important since fire is the dominant disturbance within the study area. However, these metrics may greatly benefit kNN imputation models by distinguishing pixels with similar spectral information [19,63].

5. Conclusions

The kNN imputation method is increasingly used to combine remote sensing data with ground sample plots to produce spatially explicit predictions of forest biomass at the landscape scale. While studies have demonstrated different kNN imputation approaches for estimating forest biomass, there is currently no consensus on which method is most appropriate when integrating multispectral time-series with field inventory data. This study addresses this gap by comparing different kNN distance techniques (with k = 1) and biomass imputation methods (direct and indirect). We found that the best results of forest biomass predictions can be achieved using the indirect imputation method rather than the direct method. In addition, our results confirm that RF outperforms GNN and MSN distance techniques in biomass imputation models. Our recommendation is that land managers and researchers should consider using a RF-based kNN imputation model that incorporates

Landsat-based time-series metrics with forest structure variables (basal area and stem density) for estimating forest biomass.

Author Contributions: Conceptualization, H.T.N., S.J., M.S.-B., and A.H.; Formal analysis, H.T.N. and S.H.; Methodology, H.T.N. and S.J.; Software, H.T.N.; Supervision, S.J., A.H. and M.S.-B.; Writing—original draft, H.T.N.; Writing—review & editing, S.J., M.S.-B., S.H. and A.H.

Funding: This research received no external funding.

Acknowledgments: This research was partly funded by the Australian Award Scholarship and the Cooperative Research Centre for Spatial Information (CRCSI) under Project 4.104 (A Monitoring and Forecasting Framework for the Sustainable Management of southeast Australian Forests at the Large Area Scale). CRCSI activities are funded by the Australian Commonwealth's Cooperative Research Centres Programme. We also would like to acknowledge the work of Liam Costello from the Department of Environment, Land Water and Planning (DELWP) on extracting variables from VFMP dataset.

Conflicts of Interest: The authors declare no conflict of interest. The funding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; or the writing of the manuscript.

References

1. Houghton, R.A.; Hall, F.; Goetz, S.J. Importance of biomass in the global carbon cycle. *J. Geophys. Res. Biogeosci.* **2009**, *114*. [CrossRef]
2. Tomppo, E.; Olsson, H.; Ståhl, G.; Nilsson, M.; Hagner, O.; Katila, M. Combining national forest inventory field plots and remote sensing data for forest databases. *Remote Sens. Environ.* **2008**, *112*, 1982–1999. [CrossRef]
3. Wulder, M.; Skakun, R.; Kurz, W.; White, J. Estimating time since forest harvest using segmented Landsat ETM+ imagery. *Remote Sens. Environ.* **2004**, *93*, 179–187. [CrossRef]
4. Cao, L.; Coops, N.C.; Innes, J.L.; Sheppard, S.R.J.; Fu, L.; Ruan, H.; She, G. Estimation of forest biomass dynamics in subtropical forests using multi-temporal airborne LiDAR data. *Remote Sens. Environ.* **2016**, *178*, 158–171. [CrossRef]
5. Badreldin, N.; Sanchez-Azofeifa, A. Estimating Forest Biomass Dynamics by Integrating Multi-Temporal Landsat Satellite Images with Ground and Airborne LiDAR Data in the Coal Valley Mine, Alberta, Canada. *Remote Sens.* **2015**, *7*, 2832–2849. [CrossRef]
6. Zald, H.S.J.; Ohmann, J.L.; Roberts, H.M.; Gregory, M.J.; Henderson, E.B.; McGaughey, R.J.; Braaten, J. Influence of lidar, Landsat imagery, disturbance history, plot location accuracy, and plot size on accuracy of imputation maps of forest composition and structure. *Remote Sens. Environ.* **2014**, *143*, 26–38. [CrossRef]
7. Meyer, V.; Saatchi, S.S.; Chave, J.; Dalling, J.W.; Bohlman, S.; Fricker, G.A.; Robinson, C.; Neumann, M.; Hubbell, S. Detecting tropical forest biomass dynamics from repeated airborne lidar measurements. *Biogeosciences* **2013**, *10*, 5421–5438. [CrossRef]
8. Tsui, O.W.; Coops, N.C.; Wulder, M.A.; Marshall, P.L.; McCardle, A. Using multi-frequency radar and discrete-return LiDAR measurements to estimate above-ground biomass and biomass components in a coastal temperate forest. *ISPRS J. Photogramm. Remote Sens.* **2012**, *69*, 121–133. [CrossRef]
9. Pflugmacher, D.; Cohen, W.B.; Kennedy, R.E. Using Landsat-derived disturbance history (1972–2010) to predict current forest structure. *Remote Sens. Environ.* **2012**, *122*, 146–165. [CrossRef]
10. Waser, L.; Ginzler, C.; Rehush, N. Wall-to-Wall Tree Type Mapping from Countrywide Airborne Remote Sensing Surveys. *Remote Sens.* **2017**, *9*. [CrossRef]
11. He, Q.; Chen, E.; An, R.; Li, Y. Above-Ground Biomass and Biomass Components Estimation Using LiDAR Data in a Coniferous Forest. *Forests* **2013**, *4*, 984–1002. [CrossRef]
12. Ioki, K.; Tsuyuki, S.; Hirata, Y.; Phua, M.-H.; Wong, W.V.C.; Ling, Z.-Y.; Saito, H.; Takao, G. Estimating above-ground biomass of tropical rainforest of different degradation levels in Northern Borneo using airborne LiDAR. *Forest Ecol. Manag.* **2014**, *328*, 335–341. [CrossRef]
13. Wulder, M.A.; White, J.C.; Bater, C.W.; Coops, N.C.; Hopkinson, C.; Chen, G. Lidar plots—A new large-area data collection option: context, concepts, and case study. *Can. J. Remote Sens.* **2014**, *38*, 600–618. [CrossRef]
14. White, J.C.; Coops, N.C.; Wulder, M.A.; Vastaranta, M.; Hilker, T.; Tompalski, P. Remote Sensing Technologies for Enhancing Forest Inventories: A Review. *Can. J. Remote Sens.* **2016**, *42*, 619–641. [CrossRef]

15. Matasci, G.; Hermosilla, T.; Wulder, M.A.; White, J.C.; Coops, N.C.; Hobart, G.W.; Zald, H.S.J. Large-area mapping of Canadian boreal forest cover, height, biomass and other structural attributes using Landsat composites and lidar plots. *Remote Sens. Environ.* **2018**, *209*, 90–106. [CrossRef]
16. Bolton, D.K.; White, J.C.; Wulder, M.A.; Coops, N.C.; Hermosilla, T.; Yuan, X. Updating stand-level forest inventories using airborne laser scanning and Landsat time series data. *Int. J. Appl. Earth Observ. Geoinf.* **2018**, *66*, 174–183. [CrossRef]
17. Jiménez, E.; Vega, J.A.; Fernández-Alonso, J.M.; Vega-Nieva, D.; Ortiz, L.; López-Serrano, P.M.; López-Sánchez, C.A. Estimation of aboveground forest biomass in Galicia (NW Spain) by the combined use of LiDAR, LANDSAT ETM+ and National Forest Inventory data. *iForest Biogeosci. For.* **2017**, *10*, 590–596. [CrossRef]
18. Deo, R.; Russell, M.; Domke, G.; Andersen, H.-E.; Cohen, W.; Woodall, C. Evaluating Site-Specific and Generic Spatial Models of Aboveground Forest Biomass Based on Landsat Time-Series and LiDAR Strip Samples in the Eastern USA. *Remote Sens.* **2017**, *9*, 598. [CrossRef]
19. Zald, H.S.J.; Wulder, M.A.; White, J.C.; Hilker, T.; Hermosilla, T.; Hobart, G.W.; Coops, N.C. Integrating Landsat pixel composites and change metrics with lidar plots to predictively map forest structure and aboveground biomass in Saskatchewan, Canada. *Remote Sens. Environ.* **2016**, *176*, 188–201. [CrossRef]
20. Pflugmacher, D.; Cohen, W.B.; Kennedy, R.E.; Yang, Z. Using Landsat-derived disturbance and recovery history and lidar to map forest biomass dynamics. *Remote Sens. Environ.* **2014**, *151*, 124–137. [CrossRef]
21. Cohen, W.B.; Goward, S.N. Landsat's role in ecological applications of remote sensing. *Bioscience* **2004**, *54*, 535–545. [CrossRef]
22. Cohen, W.; Healey, S.; Yang, Z.; Stehman, S.; Brewer, C.; Brooks, E.; Gorelick, N.; Huang, C.; Hughes, M.; Kennedy, R.; et al. How Similar Are Forest Disturbance Maps Derived from Different Landsat Time Series Algorithms? *Forests* **2017**, *8*, 98. [CrossRef]
23. Nguyen, T.H.; Jones, S.D.; Soto-Berelov, M.; Haywood, A.; Hislop, S. A spatial and temporal analysis of forest dynamics using Landsat time-series. *Remote Sens. Environ.* **2018**, *217*, 461–475. [CrossRef]
24. Kennedy, R.E.; Ohmann, J.; Gregory, M.; Roberts, H.; Yang, Z.; Bell, D.M.; Kane, V.; Hughes, M.J.; Cohen, W.B.; Powell, S.; et al. An empirical, integrated forest biomass monitoring system. *Environ. Res. Lett.* **2018**, *13*, 025004. [CrossRef]
25. Gómez, C.; White, J.C.; Wulder, M.A.; Alejandro, P. Historical forest biomass dynamics modelled with Landsat spectral trajectories. *ISPRS J. Photogramm. Remote Sens.* **2014**, *93*, 14–28. [CrossRef]
26. Powell, S.L.; Cohen, W.B.; Kennedy, R.E.; Healey, S.P.; Huang, C. Observation of Trends in Biomass Loss as a Result of Disturbance in the Conterminous U.S.: 1986–2004. *Ecosystems* **2013**, *17*, 142–157. [CrossRef]
27. Main-Knorn, M.; Cohen, W.B.; Kennedy, R.E.; Grodzki, W.; Pflugmacher, D.; Griffiths, P.; Hostert, P. Monitoring coniferous forest biomass change using a Landsat trajectory-based approach. *Remote Sens. Environ.* **2013**, *139*, 277–290. [CrossRef]
28. Zhu, X.; Liu, D. Improving forest aboveground biomass estimation using seasonal Landsat NDVI time-series. *ISPRS J. Photogramm. Remote Sens.* **2015**, *102*, 222–231. [CrossRef]
29. Ohmann, J.L.; Gregory, M.J.; Roberts, H.M. Scale considerations for integrating forest inventory plot data and satellite image data for regional forest mapping. *Remote Sens. Environ.* **2014**, *151*, 3–15. [CrossRef]
30. Beaudoin, A.; Bernier, P.Y.; Guindon, L.; Villemaire, P.; Guo, X.J.; Stinson, G.; Bergeron, T.; Magnussen, S.; Hall, R.J. Mapping attributes of Canada's forests at moderate resolution through kNN and MODIS imagery. *Can. J. Forest Res.* **2014**, *44*, 521–532. [CrossRef]
31. Hudak, A.T.; Crookston, N.L.; Evans, J.S.; Hall, D.E.; Falkowski, M.J. Nearest neighbor imputation of species-level, plot-scale forest structure attributes from LiDAR data. *Remote Sens. Environ.* **2008**, *112*, 2232–2245. [CrossRef]
32. Eskelson, B.N.I.; Temesgen, H.; Lemay, V.; Barrett, T.M.; Crookston, N.L.; Hudak, A.T. The roles of nearest neighbor methods in imputing missing data in forest inventory and monitoring databases. *Scand. J. Forest Res.* **2009**, *24*, 235–246. [CrossRef]
33. Moeur, M.; Stage, A.R. Most similar neighbor: an improved sampling inference procedure for natural resource planning. *Forest science* **1995**, *41*, 337–359.
34. Ohmann, J.L.; Gregory, M.J. Predictive mapping of forest composition and structure with direct gradient analysis and nearest- neighbor imputation in coastal Oregon, U.S.A. *Can. J. Forest Res.* **2002**, *32*, 725–741. [CrossRef]

35. Liaw, A.; Wiener, M. Classification and regression by randomForest. *R news* **2002**, *2*, 18–22.
36. Chirici, G.; Mura, M.; McInerney, D.; Py, N.; Tomppo, E.O.; Waser, L.T.; Travaglini, D.; McRoberts, R.E. A meta-analysis and review of the literature on the k-Nearest Neighbors technique for forestry applications that use remotely sensed data. *Remote Sens. Environ.* **2016**, *176*, 282–294. [CrossRef]
37. Aguirre-Salado, C.A.; Treviño-Garza, E.J.; Aguirre-Calderón, O.A.; Jiménez-Pérez, J.; González-Tagle, M.A.; Valdéz-Lazalde, J.R.; Sánchez-Díaz, G.; Haapanen, R.; Aguirre-Salado, A.I.; Miranda-Aragón, L. Mapping aboveground biomass by integrating geospatial and forest inventory data through a k-nearest neighbor strategy in North Central Mexico. *J. Arid Land* **2013**, *6*, 80–96. [CrossRef]
38. Powell, S.L.; Cohen, W.B.; Healey, S.P.; Kennedy, R.E.; Moisen, G.G.; Pierce, K.B.; Ohmann, J.L. Quantification of live aboveground forest biomass dynamics with Landsat time-series and field inventory data: A comparison of empirical modeling approaches. *Remote Sens. Environ.* **2010**, *114*, 1053–1068. [CrossRef]
39. Hudak, A.T.; Strand, E.K.; Vierling, L.A.; Byrne, J.C.; Eitel, J.U.H.; Martinuzzi, S.; Falkowski, M.J. Quantifying aboveground forest carbon pools and fluxes from repeat LiDAR surveys. *Remote Sens. Environ.* **2012**, *123*, 25–40. [CrossRef]
40. Deo, R.K.; Russell, M.B.; Domke, G.M.; Woodall, C.W.; Falkowski, M.J.; Cohen, W.B. Using Landsat Time-Series and LiDAR to Inform Aboveground Forest Biomass Baselines in Northern Minnesota, USA. *Can. J. Remote Sens.* **2017**, *43*, 28–47. [CrossRef]
41. Department of Environment and Primary Industries. *Victoria's State of the Forest Report 2013*; Victorian Government: Melbourne, Australia, 2013.
42. Viridans. Victorian Ecosystems and Vegetation. Available online: http://www.viridans.com/ECOVEG/ (accessed on 27 August 2018).
43. Haywood, A.; Mellor, A.; Stone, C. A strategic forest inventory for public land in Victoria, Australia. *Forest Ecol. Manag.* **2016**, *367*, 86–96. [CrossRef]
44. Haywood, A.; Stone, C. Estimating Large Area Forest Carbon Stocks—A Pragmatic Design Based Strategy. *Forests* **2017**, *8*, 99. [CrossRef]
45. Kieth, H.; Barrett, D.; Keenan, R. *Review of Allometric Relationships for Estimating Woody Biomass for New South Wales, the Australian Capital Territory, Victoria, Tasmania and South Australia*; Australian Greenhouse Office: Canberra, Australia, 2000.
46. Key, C.; Benson, N. *Landscape Assessment: Remote Sensing of Severity, the Normalized Burn Ratio and Ground Measure of Severity, the Composite Burn Index*; FIREMON: Fire effects monitoring and inventory system; USDA Forest Service, Rocky Mountain Research Station: Ogden, UT, USA, 2005.
47. Crist, E.P. A TM tasseled cap equivalent transformation for reflectance factor data. *Remote Sens. Environ.* **1985**, *17*, 301–306. [CrossRef]
48. Duane, M.V.; Cohen, W.B.; Campbell, J.L.; Hudiburg, T.; Turner, D.P.; Weyermann, D.L. Implications of alternative field-sampling designs on Landsat-based mapping of stand age and carbon stocks in Oregon forests. *Forest Sci.* **2010**, *56*, 405–416.
49. Kennedy, R.E.; Yang, Z.; Cohen, W.B. Detecting trends in forest disturbance and recovery using yearly Landsat time series: 1. LandTrendr — Temporal segmentation algorithms. *Remote Sens. Environ.* **2010**, *114*, 2897–2910. [CrossRef]
50. Kennedy, R.E.; Yang, Z.; Cohen, W.B.; Pfaff, E.; Braaten, J.; Nelson, P. Spatial and temporal patterns of forest disturbance and regrowth within the area of the Northwest Forest Plan. *Remote Sens. Environ.* **2012**, *122*, 117–133. [CrossRef]
51. Cohen, W.B.; Yang, Z.; Kennedy, R. Detecting trends in forest disturbance and recovery using yearly Landsat time series: 2. TimeSync — Tools for calibration and validation. *Remote Sens. Environ.* **2010**, *114*, 2911–2924. [CrossRef]
52. Meigs, G.W.; Kennedy, R.E.; Cohen, W.B. A Landsat time series approach to characterize bark beetle and defoliator impacts on tree mortality and surface fuels in conifer forests. *Remote Sens. Environ.* **2011**, *115*, 3707–3718. [CrossRef]
53. Hislop, S.; Jones, S.; Soto-Berelov, M.; Skidmore, A.; Haywood, A.; Nguyen, T. Using Landsat Spectral Indices in Time-Series to Assess Wildfire Disturbance and Recovery. *Remote Sens.* **2018**, *10*, 460. [CrossRef]
54. Gallant, J.C.; Dowling, T.I.; Read, A.M.; Wilson, N.; Tickler, P.; Inskeep, C. *Second SRTM Derived Digital Elevation Models User Guide*; Geoscience Australia: Canberra, Australia, 2010.

55. Fick, S.E.; Hijmans, R.J. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *Int. J. Climatol.* **2017**, *37*, 4302–4315. [CrossRef]
56. Efron, B.; Hastie, T.; Johnstone, I.; Tibshirani, R. Least angle regression. *Ann. Statist.* **2004**, *32*, 407–499. [CrossRef]
57. Crookston, N.L.; Finley, A.O. yaImpute: an R package for kNN imputation. *J. Stat. Softw.* **2008**, *23*, 1–16. [CrossRef]
58. Gorard, S. Revisiting a 90-year-old debate: the advantages of the mean deviation. *Br. J. Educ. Stud.* **2005**, *53*, 417–430. [CrossRef]
59. Soto-Berelov, M.; Haywood, A.; Jones, S.D.; Hislop, S.; Nguyen, H.T. Creating robust reference (training) datasets for large area time series disturbance attribution. In *Remote Sensing: Time Series Image Processing*; Weng, Q.E., Ed.; Taylor and Francis: Abingdon-on-Thames, UK, 2018.
60. Nguyen, H.-T.; Soto-Berelov, M.; Jones, S.D.; Haywood, A.; Hislop, S. Mapping forest disturbance and recovery for forest dynamics over large areas using Landsat time-series remote sensing. *Proc. SPIE* **2017**, 10421. [CrossRef]
61. Bartels, S.F.; Chen, H.Y.H.; Wulder, M.A.; White, J.C. Trends in post-disturbance recovery rates of Canada's forests following wildfire and harvest. *Forest Ecol. Manag.* **2016**, *361*, 194–207. [CrossRef]
62. Ohmann, J.L.; Gregory, M.J.; Roberts, H.M.; Cohen, W.B.; Kennedy, R.E.; Yang, Z. Mapping change of older forest with nearest-neighbor imputation and Landsat time-series. *Forest Ecol. Manag.* **2012**, *272*, 13–25. [CrossRef]
63. Matasci, G.; Hermosilla, T.; Wulder, M.A.; White, J.C.; Coops, N.C.; Hobart, G.W.; Bolton, D.K.; Tompalski, P.; Bater, C.W. Three decades of forest structural dynamics over Canada's forested ecosystems using Landsat time-series and lidar plots. *Remote Sens. Environ.* **2018**, *216*, 697–714. [CrossRef]
64. Gschwantner, T.; Lawrence, M.; McRoberts, R.E. (Eds.) *National Forest Inventories*; Springer: Dordrecht, The Netherlands, 2009.
65. Gagliasso, D.; Hummel, S.; Temesgen, H. A Comparison of Selected Parametric and Non-Parametric Imputation Methods for Estimating Forest Biomass and Basal Area. *Open J. For.* **2014**, *04*, 42–48. [CrossRef]