Remote Sensing of Environment 325 (2025) 114758

Contents lists available at ScienceDirect

Remote Sensing of Environment

journal homepage: www.elsevier.com/locate/rse

Uncertainty quantification for forest attribute maps with conformal prediction and *k*-nearest neighbor method

M. Kuronen [a], J. Räty [b], P. Packalen [a], M. Myllymäki [a,*]

[a] Natural Resources Institute Finland (Luke), Latokartanonkaari 9, FI 00790, Helsinki, Finland
[b] Natural Resources Institute Finland (Luke), Yliopistokatu 6B, FI 80100, Joensuu, Finland

ARTICLE INFO

Edited by Marie Weiss

Keywords:
Airborne laser scanning
Coverage
Jackknife
Prediction interval
Quantile regression
Remote sensing
Satellite data

ABSTRACT

Forest attribute maps relying on remotely sensed data are increasingly required for local decision-making related to the use of forest resources. Such maps always have uncertainty, which can be challenging to quantify. The objective of this work is to introduce the conformal prediction methodology to uncertainty quantification in forest attribute mapping, particularly for the *k*-NN method. We compare several conformal *k*-NN procedures for the mapping of total volume, broadleaved volume and Lorey's height using Sentinel-2 satellite images and airborne laser scanning data. We show that all procedures produce valid prediction intervals in the sense that they contain the true value with the desired probability, for example 90%. We use multiple measures to quantify how well the prediction intervals adapt to the difficulty of prediction in different forest strata. We found that there are multiple methods for *k*-NN to produce prediction intervals competitive with those produced by conformal quantile regression. These methods include conformal prediction based on the standard deviation or quantiles of the *k* nearest neighbors with commonly used values of *k*. We present how to produce a forest attribute map with the proposed conformal prediction intervals. We also show a theoretical coverage guarantee for the jackknife conformal *k*-NN procedure. We recommend conformal prediction for unit-level uncertainty quantification of forest attribute maps.

1. Introduction

Remote sensing technologies have enabled production of spatially explicit estimates of forest attributes in the form of a map (McRoberts et al., 2010; Kangas et al., 2018; Fassnacht et al., 2024). Maps are needed particularly for local decision making, e.g., in the process of planning the use of forests, but inaccuracies in the maps can lead to suboptimal decisions (Kangas et al., 2023). In practice, large differences have been observed in maps produced for the same purposes (Zhang et al., 2019; Schulp et al., 2014; Mitchard et al., 2013). Some maps have been heavily criticized due to their misleading content and contradiction with national forest inventory data based on field measurements (Palahí et al., 2021; Breidenbach et al., 2022). These types of observations have led to the call for local and global accuracy measures (Meyer and Pebesma, 2021, 2022) and methods to address the map uncertainty (Kangas et al., 2023). Kangas et al. (2018) also emphasized the importance to equip map products with valid uncertainties to understand if they are sufficiently accurate for decision-making. While uncertainties are required at different scales, from unit level (i.e., grid cell or pixel level) to small-area and even population level (e.g., McRoberts and Tomppo, 2007; McRoberts et al., 2011), this work is concerned with uncertainty quantification at the unit level.

The *k*-nearest neighbor (*k*-NN) method has been studied extensively in the context of forest resource mapping (see Chirici et al., 2016; McRoberts et al., 2011, 2010, and references therein). It is used, e.g., for national forest attribute maps (Mäkisara et al., 2022; Tomppo and Halme, 2004) and remote sensing based stand-level forest management inventories (Maltamo and Packalen, 2014). Overall, the number of countries with *k*-NN applications is large (McRoberts et al., 2011).

Several approaches have been suggested in the literature to estimate uncertainty for *k*-NN predictions. Leave-one-out RMSE has been used conventionally as a unit level uncertainty measure for the *k*-NN predictions (Kim and Tomppo, 2006). These uncertainty metrics are reported, e.g., for the Finnish multi-source national forest inventory maps (Mäkisara et al., 2022). Leave-one-out RMSEs are also common in studies focusing on stand-level management inventories (Kotivuori et al., 2021). However, RMSE is not a location specific uncertainty measure: it essentially assumes the same amount of uncertainty everywhere in the area.

* Corresponding author.
E-mail address: mari.myllymaki@luke.fi (M. Myllymäki).

https://doi.org/10.1016/j.rse.2025.114758
Received 26 November 2024; Received in revised form 2 April 2025; Accepted 11 April 2025
Available online 1 May 2025

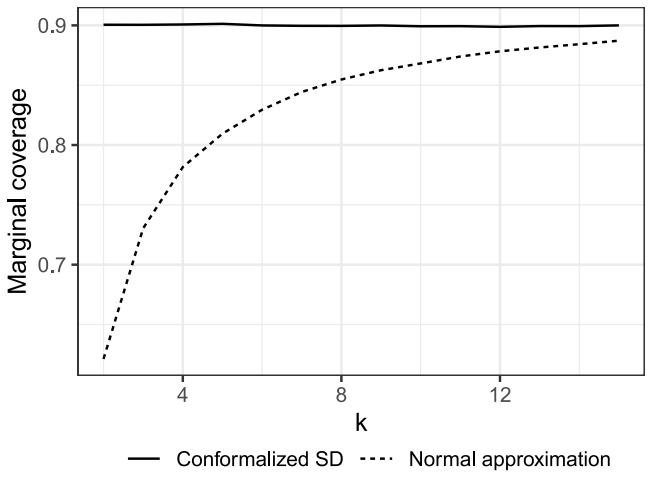

Fig. 1. Marginal coverage for the naive interval based on normal approximation and standard deviation (SD) estimated from k nearest neighbors (dashed line), and the conformalized interval based on the same SD (solid line). The target coverage was 0.90. (The variable was total volume in the Central Finland study area shown in Fig. 3.).

In the desire to provide location specific uncertainty measure, other approaches than RMSE must be used. Kim and Tomppo (2006) proposed a location specific method for the unit level uncertainty estimation based on fitting a variogram on the feature space. Sagar et al. (2022) proposed a method in which uncertainty is estimated from the collection of maps produced with training data obtained by bootstrapping. Sagar et al. (2022) also proposed using the mean distance to neighbors to investigate the area of applicability of the model, given the training data. Further, there are some intuitively appealing uncertainty estimates like for example the standard deviation (SD) of the k nearest neighbors.

The produced uncertainty intervals should preferably contain the ground truth with a user-specified probability, such as 90%. This property was not examined in any of the articles mentioned above, and there is generally no guarantee that the desired 90% coverage is obtained. For example, using the normal approximation based on the SD of the k nearest neighbors can lead to severe undercoverage (Fig. 1, dashed line). Thus, there is a clear need to enhance the uncertainty quantification related to the unit level k-NN predictions.

Conformal prediction is a user-friendly method to make proper uncertainty estimates out of heuristic ones, guaranteeing finite sample coverage (e.g., Angelopoulos and Bates, 2022). It is a general method which can be used with any prediction method. For example, the standard deviation of the k nearest neighbors based intervals can be corrected to have the 90% coverage with conformalization (Fig. 1, solid line). The general idea is to shrink or extend the heuristic prediction intervals such that the desired coverage is obtained, as illustrated in Fig. 2 (left) for an interesting variable y with respect to a single input x. The amount of adjustment needed for the prediction intervals is derived from a calibration data set.

In general, there are several alternatives for how to transform heuristic uncertainty intervals to conformal prediction intervals (Angelopoulos and Bates, 2022). The split method is the most widely used version of conformal prediction and generally applicable (Angelopoulos and Bates, 2022). In this method, a part of the training data is separated from the rest to be used for calibration. However, the jackknife conformal prediction (Vovk, 2015; Barber et al., 2021) is particularly attractive for the k-NN method, because it can be efficiently implemented together with the jackknife mean k-NN predictions. The jackknife method can generally be preferred because it allows to use all the training data, without the need to leave part of the training data to a separate calibration set. However, it is typically computationally too intensive, as it requires fitting the model n times (n being the number of data points). Also its theoretical coverage properties are not known in a general case.

In the machine learning literature, Papadopoulos et al. (2011) proposed prediction intervals for the k-NN method based on conformal prediction. However, they worked with the split method. On the other hand, Barber et al. (2021) showed that the jackknife method works for the k-NN method, but only with simple scores that produce prediction intervals of constant width. Constant width of the prediction intervals means that while the intervals achieve the 90% coverage, they are not adapted to heteroscedasticity, which forest attributes often tend to have. Fig. 2 (right) illustrates the problem: constant prediction intervals are far too wide for small x and too narrow for large x. Thus, we generalize the results of Barber et al. (2021) for the use of k-NN with heuristic uncertainty measures derived from the k nearest neighbors and investigate alternatives for the heuristic measure.

There are a few examples where conformal prediction was used in the context of remote sensing of environment. For example, Singh et al. (2024) illustrated conformal prediction for land cover classification, tree canopy height estimation and detection of invasive tree species. Kakhani et al. (2024) compared different quantile regression methods and conformal prediction for providing uncertainty quantification for soil organic carbon estimation. Further, conformal classification was used for classifying, e.g., forest health status (Norinder and Lowry, 2023) and land-use land-cover (Valle et al., 2023). In all these studies' regression cases, conformal quantile regression and the split method were used.

The aim of the paper is to introduce the conformal prediction to forest attribute mapping using different types of remotely sensed data and the popular k-NN method. First, we show that the jackknife conformal prediction can be used with the k-NN method instead of the split conformal prediction. Second, we evaluate the conformal k-NN methods based on various heuristic uncertainty measures in the forest attribute mapping case using multiple efficiency measures found from the literature. The efficiency measures aim to quantify how well the prediction intervals adjust to the heteroscedasticity of data (see Fig. 2). We use the conformalized quantile regression as a reference method since according to Angelopoulos and Bates (2022) it is often the best way to obtain continuous prediction intervals. Third, we give step-by-step guidance how to produce forest attribute maps with conformal uncertainty intervals using remote sensing data and field samples (see Section 3.5).

2. Data

We used Copernicus Sentinel-2 (S2) images and ALS data in the study. A mosaic of the S2 images covered whole Finland excluding the northern Lapland, whereas the ALS data were only utilized in a 1.7 Mha study area located in Central Finland (Fig. 3). National forest inventory (NFI) plots of Finland were used as field samples (Korhonen et al., 2021). The analyses were restricted to the forest area defined by the forest mask from MS-NFI (Mäkisara et al., 2022). More detailed description of the field and remotely sensed data follows.

2.1. Field data

We used the Finnish NFI data (Korhonen et al., 2024) from years 2019–2021 in the whole Finland study area, whereas the NFI data of years 2019–2023 were used in the Central Finland study area. The Finnish NFI is based on a systematic cluster sampling and circular concentric plots. Each cluster contains 8–11 plots depending on the sampling regions defined in Korhonen et al. (2024). Trees with diameter at breast height (dbh) $\geq$ 9.5 cm are measured using a radius of 9 m, trees with 4.5 cm $\leq$ dbh $<$ 9.5 cm using a radius of 4 m, and smaller trees with a relascope with factor 1.5. Heights of non-sample trees and upper diameter at 6 m for all trees were predicted

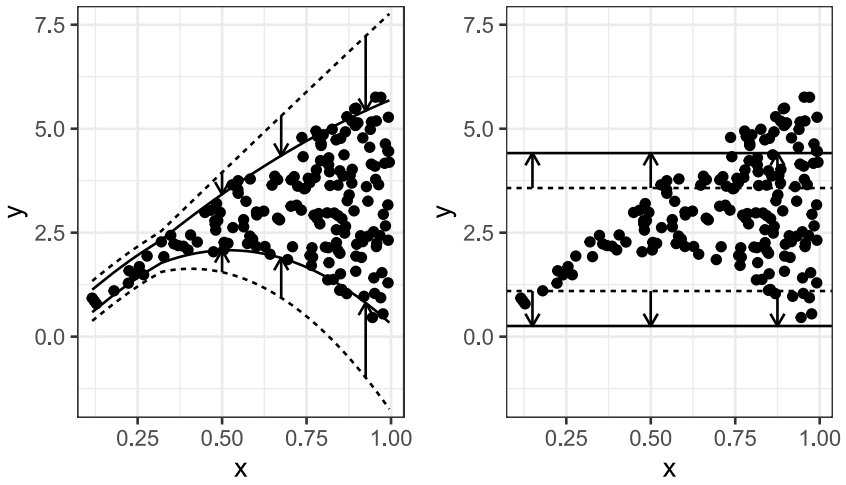

Fig. 2. Illustration of conformal prediction. Dashed lines represent heuristic prediction intervals for variable y with coverage above (left) or below (right) the desired 90%. Conformal prediction adjusts the heuristic intervals to reach 90% coverage (solid lines). Arrows relate to the adjustment.

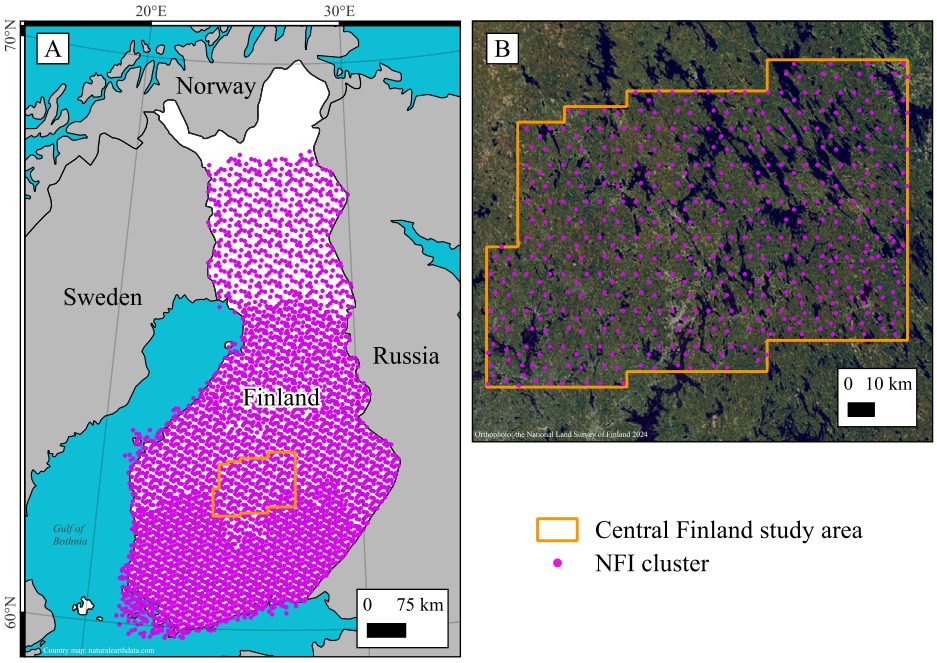

Fig. 3. (A) The National forest inventory (NFI) clusters of the study area covering whole Finland excluding the northernmost part of the country and (B) the NFI clusters of the Central Finland study area.

as described in Korhonen et al. (2024). Stem volume for each tree was predicted using the species-specific three-predictor volume models proposed by Laasasenaho (1982). In this work, we used the total volume (m^3/ha), broadleaf volume (m^3/ha) and Lorey's height (m) computed from the tree level data for each field plot.

As we were interested in mapping the forest attributes in the forest mask only, we utilized only the field plots located within the mask. There were 30433 field plots in the whole country and 2734 field plots in the Central Finland study area.

2.2. Remote sensing data

The mosaicking of the S2 images was carried out as a manual work as explained in Pitkänen et al., 2024. The mosaic is based on S2 Level-2 A images that were downloaded from the Copernicus Open Access Hub. In the Level-2 A images the pixel values represent the bottom-of-atmosphere reflectance and images are orthorectified. The S2 images with as low cloud coverage as possible were selected from a time window of early June and late August 2020–2021 for the creation of the image mosaic.

ALS data of the Central Finland study area was acquired in seven projects between 2019 and 2023 (Table 1). These acquisitions are part of the national data collection campaign (KALLIO) organized by the National Land Survey of Finland. Data acquisition parameters and sensors varied across projects. We used harmonized ALS data having the nominal pulse density of 0.5 pulses/m^2. This data is publicly available through OGC API - Processes interface (National Land Survey of Finland, 2024). The ALS echoes were classified as ground and other returns (Axelsson, 2000), and a digital terrain model was interpolated based on the ground returns. The ALS echoes were height-normalized

Table 1
Acquisition parameters of airborne laser scanning data in the Central Finland study area. Coverage refers to the areal coverage of the acquisition in the study area.

Year	Sensor	Leaf condition	Flight altitude (m)	Coverage (%)
2023	Riegl VQ-780IIS	Leaf-on	1400	17.4
2022	Riegl VQ-780IIS	Leaf-on	1445	17.4
2021	Riegl VQ-780II	Leaf-on	1340	17.4
2020	Riegl VQ-780i	Leaf-on	1280	17.0
2020	Leica ALS70HA	Leaf-off	2621	1.3
2019	Riegl VQ-1560i	Leaf-on	1704	17.4
2019	Leica ALS80HP	Leaf-on	2837	12.0

with respect to the ground level by subtracting the digital terrain model from the original echo heights.

2.3. Feature variables

We computed feature variables from the echo heights of the ALS data and from S2 images for the NFI field plots and grid cells of size 16 m×16 m. ALS features were computed without a height cutoff from the combined set of first-of-many and only echoes. The calculated ALS features were means, medians, height quantiles, standard deviations and densities. Height quantiles (5, 10, 15, …, 100%) were calculated using the default method in R (R Core Team, 2023; Hyndman and Fan, 1996) and densities were computed by dividing the number of echoes below a height threshold (2, 5, 10 and 15 m) by the total number of echoes. S2 features were computed as area-weighted averages (Baston, 2023).

3. Methods

3.1. The k-nearest neighbors (k-NN) method

We use the following terminology: the remote sensing variables or possible other ancillary variables are feature variables (or features), the space defined by the feature variables is the feature space, the set of all population units for which observations of both response and feature variables are available is the training data (also called sometimes reference set), and the set of population units for which estimates are required is the target set. The estimate for the ith target unit is calculated in the k-NN method as

$$\hat{y}_i = \sum_{j=1}^{k} w_{i_j} y_{i_j}, \tag{1}$$

where i_j refers to the jth nearest neighbor of unit i from the training data, y_{i_j} is the observed value of the response variable for this jth nearest neighbor and w_{i_j} is the associated weight with the property $\sum_{j=1}^{k} w_{i_j} = 1$.

The k nearest neighbors are specified by a distance metric in the feature space, i.e., the k units j with the smallest distance d_{ij} from the unit i are chosen as the neighbors. Numerous distance metrics, such as Euclidean and most similar neighbor, have been suggested for the prediction of forest attributes (Packalén et al., 2012; Cosenza et al., 2021). After some preliminary runs, we decided to use the Euclidean distance metric to determine the nearest neighbors in the feature space. Prior to computing distances, all the feature variables were standardized to put them on the same scale by subtracting the mean and dividing by the standard deviation.

In the case of a continuous response variable the predicted value is an average or weighted average of the response variable of the k nearest neighbors. Commonly used weighting scheme is inverse distance such that $w_{i_j} \propto d_{ij}^{-t}$ and $0 \le t \le 2$. If $t = 0$, then equal weights are assigned to all k nearest neighbors. Initially we tested the options $t = 0$ and 1, but as the k-NN predictions (Eq. (1)) were little affected by the choice of t, we chose the simpler alternative $t = 0$.

To produce a map, in a typical setup, a grid is overlaid on the study region. The grid has a large number of grid cells, G, and the predictions are required for each grid cell. This collection of grid cells forms then the target set.

For the data from whole Finland, our feature variables included 10 S2 features and we used all of these in the k-NN. For the data from Central Finland, we extracted 30 ALS features and 10 S2 features that were used as predictor variable candidates for the k-NN models. The selection of the predictor variables was carried out by using the simulated annealing algorithm (Kirkpatrick et al., 1983). As a variable selection tool, the simulated annealing can be used to iteratively search for the optimal set of predictor variables (Packalén et al., 2012). Our variant of the simulated annealing was controlled by an initial temperature and a cooling factor and a number of inner iterations per temperature that defined the ultimate number of iterations. The initial temperature was set to 1, the cooling factor was 0.95, and the algorithm was iterated 10 times per temperature. The algorithm used mean squared error as a cost function. For each k-NN model, we selected 5 predictor variables.

For volume, the selected ALS features were the mean and 90% quantile of ALS echo heights and the means of spectral recordings by S2 bands B3 (green), B7 (red edge), and B8 (near-infrared). For broadleaved volume, the selected ALS features were the proportion of ALS echoes returned below the 10-meter threshold and the 85% quantile of ALS echoes. Among the S2 features, the means of spectral recordings by S2 bands B2 (blue), B8 (near-infrared) and B12 (short-wave infrared) were selected as predictor variables for the broadleaved volume model. For Lorey's height, the variable selection procedure only selected ALS features. The ALS features were mean, and 55%, 75%, 95% and 100% quantiles computed from the echo heights.

3.2. Conformal prediction

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ denote the i.i.d. training data and (X_{n+1}, Y_{n+1}) a new data point drawn independently from the same distribution as the training data. Conformal prediction is an approach to produce a predictive set $\hat{C}$ which contains the new data point Y_{n+1} with the desired coverage level $1-\alpha$. Formally, the predictive set $\hat{C}$ should satisfy the following coverage property

$$P(Y_{n+1} \in \hat{C}(X_{n+1})) \approx 1 - \alpha, \tag{2}$$

where the probability is taken with respect to the new data point and the training data (Barber et al., 2021). In the following, we present the split (or inductive) conformal prediction (Section 3.2.1), the simplest score and calibration used in the conformal prediction (Section 3.2.2), adaption of the conformal prediction to heteroscedasticity by using heuristic uncertainties (Section 3.2.3) and using conformalized quantile regression (Section 3.2.4). Finally, we present the conformal jackknife prediction method, and show a theoretical coverage guarantee for it with the k-NN method.

3.2.1. Split conformal prediction

In split conformal prediction the training data is first split into the so-called proper training set $(x_i, y_i), i \in \mathcal{I}_1$, and the calibration set $(x_i, y_i), i \in \mathcal{I}_2$. The proper training set is used for training the prediction model and the calibration set is used to calibrate the prediction sets.

3.2.2. Simple conformal prediction

A conformity score

$$s_i = |y_i - \hat{f}(x_i)| \tag{3}$$

is computed for each example i in the calibration set $\mathcal{I}_2$. Here $\hat{f}(x_i)$ is the prediction for y_i given x_i and the model f. Let $\hat{q}$ be the $\lceil (1-\alpha)(n+1) \rceil / n$ quantile of the conformity scores $s_i \in \mathcal{I}_2$. Then the predictive interval for a given x is

$$C(x) = (\hat{f}(x) - \hat{q}, \hat{f}(x) + \hat{q}).$$

The problem with the basic form of conformal prediction in the regression setting is that the interval length is independent of x. If it is known that a transformation of the response leads to a more homoscedastic model, then an effective way of obtaining better prediction intervals is doing the conformal prediction on the transformed response and back transforming the intervals. An example is the logarithmic transformation

$$g(y) = \log(y + \lambda),$$

with the inverse transformation

$$g^{-1}(y) = \exp(y) - \lambda,$$

where λ is an offset. The conformity score is $s_i = |g(y_i) - \hat{f}(x_i)|$, where $\hat{f}(x_i)$ is the prediction for $g(y_i)$ given x_i and model f. The conformal prediction interval is then

$$C(x) = (g^{-1}(\hat{f}(x) - \hat{q}), g^{-1}(\hat{f}(x) + \hat{q})).$$

3.2.3. Conformal prediction with heuristic uncertainty estimates

To adapt conformal predictions to heteroscedasticity, explicitly modeled uncertainties or heuristic uncertainty estimates can be utilized (e.g. Papadopoulos et al., 2008, 2011; Lei et al., 2018). First, a heuristic (or an explicitly modeled) uncertainty $u(x_i)$ is computed for each example i in the calibration set $\mathcal{I}_2$. Second, a conformity score

$$s_i = \frac{|y_i - \hat{f}(x_i)|}{u(x_i)} \tag{4}$$

is computed for each $i \in \mathcal{I}_2$. Third, $\hat{q}$, i.e., the $\lceil (1-\alpha)(n+1) \rceil / n$ quantile of the conformity scores, is found. Finally, the predictive interval for a given x is

$$C(x) = (\hat{f}(x) - \hat{q}u(x), \hat{f}(x) + \hat{q}u(x)). \tag{5}$$

Papadopoulos et al. (2011) proposed several heuristic uncertainty measures for the k-NN method. These measures are based on the standard deviation of the k nearest neighbors,

$$sd_k(x_i) = \sqrt{\frac{1}{k}\sum_{j=1}^{k}\left(y_{i_j} - \frac{1}{k}\sum_{l=1}^{k} y_{i_l}\right)^2}, \tag{6}$$

or the sum of the distances to the k nearest neighbors,

$$d_k(x_i) = \frac{1}{k}\sum_{j=1}^{k} \mathrm{dist}(x_{i_j}, x_i). \tag{7}$$

We considered also heuristic uncertainties based on the root mean squared error of the k nearest neighbors of unit i, i.e.,

$$e(x_i) = \sqrt{\frac{1}{k}\sum_{j=1}^{k}(y_{i_j} - \hat{y}_{i_j})^2}, \tag{8}$$

which were previously utilized in the small-area estimation context by Kangas et al. (2024).

3.2.4. Conformalized quantile regression

Another somewhat different approach to conformal prediction, is the conformalized quantile regression (Romano et al., 2019), which directly aims at predicting the interval instead of point estimates. Romano et al. (2019) proposed the conformalized quantile regression within the split conformal prediction framework. It is implemented as follows: Let $\alpha_{\mathrm{lo}} = \alpha/2$ and $\alpha_{\mathrm{hi}} = 1 - \alpha/2$ be the lower and upper quantiles. Quantile regression estimates the α_{lo}th and α_{hi}th quantiles of the response variable y as a function of x. Let $\hat{q}_{\alpha_{\mathrm{lo}}}$ and $\hat{q}_{\alpha_{\mathrm{hi}}}$ be the prediction functions for the lower and upper quantiles. The conformity score is

$$s_i = \max\{\hat{q}_{\alpha_{\mathrm{lo}}}(x_i) - y_i, y_i - \hat{q}_{\alpha_{\mathrm{hi}}}(x_i)\} \tag{9}$$

for each $i \in \mathcal{I}_2$. Let $\hat{q}$ be the $\lceil (1-\alpha)(n+1) \rceil / n$ quantile of the scores. The prediction interval for a new data x is

$$C(x) = (\hat{q}_{\alpha_{\mathrm{lo}}}(x) - \hat{q}, \hat{q}_{\alpha_{\mathrm{hi}}}(x) + \hat{q}). \tag{10}$$

Conformalized quantile regression can be used with any quantile regression method. We considered random forest (quantregForest; Meinshausen, 2006) and k-NN (yaImpute; Crookston and Finley, 2007) based quantile regression. The random forest quantile regression directly targets specific quantiles. However, with conformalized quantile regression targeting a specific quantile is not necessary. Thus, for the k-NN quantile regression, we simply tested (a) the minimum and maximum values among the k nearest neighbors, i.e.,

$$\hat{q}_{\alpha_{\mathrm{lo}}}(x_i) = \min\{y_{i_j} : j = 1, \ldots, k\} \text{ and } \hat{q}_{\alpha_{\mathrm{hi}}}(x_i) = \max\{y_{i_j} : j = 1, \ldots, k\}, \tag{11}$$

and (b) the α_{lo}th and α_{hi}th sample quantiles of the values among the k nearest neighbors as the quantile predictions.

3.2.5. Conformal jackknife prediction

Vovk (2015) proposed and empirically studied the cross-conformal predictor, which is an alternative to the split conformal method. In cross-conformal predictor the training set is split into K subsets (folds) $\mathcal{J}_1, \ldots, \mathcal{J}_K$. We consider here the special case of leave-one-out or jackknife predictor with $K = n$.

Let us denote the conformity score function which does not utilize the ith point by

$$\hat{s}_{-i} = \mathcal{A}((X_1, Y_1), \ldots, (X_{i-1}, Y_{i-1}), (X_{i+1}, Y_{i+1}), \ldots, (X_n, Y_n)),$$

where $\mathcal{A}$ is a conformity score algorithm. The functions $\hat{s}_{-i}$ are used to compute $\hat{q}$, i.e., the $\lceil (1-\alpha)(n+1) \rceil / n$ quantile of the scores $s_i = \hat{s}_{-i}(x_i, y_i)$, $i = 1, \ldots, n$. Then the conformal prediction set for a new point x utilizes the conformity score function which uses all the training data,

$$\hat{s} = \mathcal{A}((X_1, Y_1), \ldots, (X_n, Y_n)).$$

That is, the conformal prediction set for x is

$$\hat{C}(x) = \{y : \hat{s}(x, y) < \hat{q}\}.$$

Theoretically much less is known about jackknife than the split method. Recently Barber et al. (2021) proved a coverage guarantee for jackknife with the simple score (Eq. (3)) under some conditions on the prediction algorithm. In Appendix A, we extend this result by showing a similar coverage guarantee for jackknife with any conformity score based on the k nearest neighbors (examples given in Sections 3.2.3 and 3.2.4). That is, the $100(1-\alpha)\%$ k-NN jackknife conformal prediction intervals based on a conformity score derived from the k nearest neighbors satisfy

$$P(Y_{n+1} \in \hat{C}(X_{n+1})) \geq 1 - \alpha - 2\sqrt{k/n}. \tag{12}$$

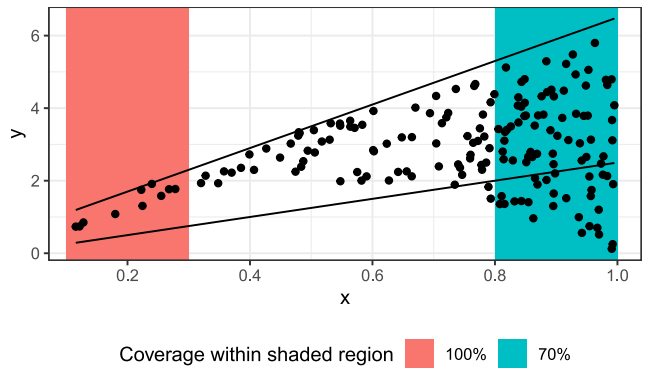

Fig. 4. The shaded regions represent two possible 20% intervals for the data sorted according to the x variable. In the left shaded region the prediction bands (black lines) cover 100% of the observations and in the right shaded region only 70%. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3.3. Evaluation measures

The different conformal procedures presented above all satisfy the marginal coverage property (Eq. (2)). However, ideally prediction intervals should have the nominal coverage for every input x. This property is called conditional coverage (Angelopoulos and Bates, 2022). Figs. 2 and 4 show examples of prediction intervals. While marginal coverages are fine in each case, there are clear differences in conditional coverages with respect to x. Although ideal conditional coverage is impossible (Vovk, 2013), good conditional coverage is a desirable property. Besides wishing good conditional coverage with respect to inputs x, in forest attribute mapping it would be important to have equal coverage in different geographic regions, for example. Imagine, for example, that the country for which volume maps are constructed is split to ten provinces with equal number of data units. In one of the provinces the predictions sets never cover the observed volume, while, in all the other provinces, the prediction sets always cover the observed volume. Then the prediction sets have 90% marginal coverage across the whole country, but not conditional coverage. Conditional coverage would imply that the prediction sets cover the observed volume at least 90% of the time in all provinces.

Many metrics have been proposed to evaluate conformal procedures. The feature-stratified coverage is defined as the minimum coverage in groups defined by a discrete variable, e.g., geographical region (Angelopoulos and Bates, 2022). A continuous variable can be discretized and then feature-stratified coverage can be computed for a continuous variable (see two strata illustrated in Fig. 4).The size-stratified coverage is a modification of feature-stratified coverage where instead of a feature the prediction interval length is used. Worst-slab coverage metric (Cauchois et al., 2021) uses random linear combinations of the features and a sliding window approach instead of a prespecified stratification.

Since, in our case all the features were continuous, we applied the sliding window approach of worst-slab coverage to the feature-stratified and size-stratified coverage metrics, to avoid manually discretizing the variables. Additionally we computed worst-slab coverage with respect to the northing and the response. For simplicity we used the northing instead of geographical regions.

Formally, worst-slab coverage is obtained by going through all the possible intervals (a, b) which cover at least $\delta \cdot 100\%$ of the data (e.g., $\delta = 0.20$) and computing the minimal coverage across all these intervals, i.e.,

$$\mathrm{WSC}(\hat{C}) = \min_{b-a>\delta n} \left\{ \frac{1}{b-a+1} \sum_{i=a}^{b} 1\left(Y_{(i)} \in \hat{C}(X_{(i)})\right) \right\}, \tag{13}$$

where $(X_{(i)}, Y_{(i)})_{i=1}^{n}$ is the data ordered according to (1) the response variable Y (WSCY), (2) northing (WSCnorth), or (3) the interval length $|\hat{C}(X)|$ (WSCsize). For the worst-slab coverage of Cauchois et al. (2021), the data was ordered according to random linear combination of the features X. This was repeated 100 times and the mean of the resulting coverages is called WSCX.

There are also many measures that do not as directly measure conditional coverage. We inspected the interval score (Gneiting and Raftery, 2007). The interval score for one unit y is defined as

$$S_\alpha(l, u; y) = (u - l) + \frac{2}{\alpha}(l - y)\mathbf{1}(y < l) + \frac{2}{\alpha}(y - u)\mathbf{1}(y > u),$$

where l and u are the lower and upper boundaries of the interval, respectively. It is the larger, the larger the interval and exceedances (below or above) of the interval are. Thus, the smaller interval score, the better. The interval score for the conformal procedure is computed as the mean of the unit-specific scores,

$$\mathrm{IScore} = \frac{1}{n} \sum_{i=1}^{n} S_\alpha(l_i, u_i; y_i). \tag{14}$$

It is proper scoring rule for predicting the quantiles at level $\alpha/2$ and $1 - \alpha/2$ (e.g., Gneiting and Raftery, 2007).

To combine all measures into one score, we had to do some normalization. As probabilities, the worst-slab coverages did not need any normalization to be comparable. However, the interval score is dependent on the scale of the variable in question and the difficulty of predicting it. To get a comparable score we divided the interval score (14) by the minimum interval score observed for the same variable, dataset and k. We further scaled this normalized interval score (IScoreN) so that twice the best interval score would result in the same penalty as having 80% worst-slab coverage. Thus, we defined the overall score as

$$\begin{aligned}\mathrm{Score} = {} & \mathrm{WSCY} + \mathrm{WSCnorth} + \mathrm{WSCsize} + \mathrm{WSCX} \\ & + (1 - 0.1(\mathrm{IScoreN} - 1)).\end{aligned} \tag{15}$$

3.4. Simulation study setup

We tested the following conformal predictions for the k-NN method:

- Simple conformal prediction using scores of Eq. (3) for the logarithmically transformed response

$$\log(y + \lambda) \tag{16}$$

  with offset $\lambda = 0, 1, 2, \ldots, 100$,
- Conformal prediction with heuristic uncertainties based on the standard deviation (Eq. (6)) of the k nearest neighbors with

$$u(x_i) = \frac{sd_k(x_i)}{\mathrm{med}_i(sd_k(x_i))} + \gamma \quad \text{and} \tag{17}$$

$$u(x_i) = \exp\left(\gamma \cdot \frac{sd_k(x_i)}{\mathrm{med}_i(sd_k(x_i))}\right), \tag{18}$$

  where $\mathrm{med}_i(sd_k(x_i))$ is the median of $sd_k(x_i)$ of all units $i \in \mathcal{I}_1 \cup \mathcal{I}_2$, as suggested by Papadopoulos et al. (2011), and $\gamma = 0, 0.1, 0.2, 0.3, \ldots, 2.0$,
- Conformal prediction with heuristic uncertainties based on the mean distance (Eq. (7)) to the k nearest neighbors with

$$u(x_i) = \frac{d_k(x_i)}{\mathrm{med}_i(d_k(x_i))} + \gamma, \quad \text{and}$$

$$u(x_i) = \exp\left(\gamma \cdot \frac{d_k(x_i)}{\mathrm{med}_i(d_k(x_i))}\right),$$

  where $\mathrm{med}_i(d_k(x_i))$ is the median of $d_k(x_i)$ of all units $i \in \mathcal{I}_1 \cup \mathcal{I}_2$, suggested by Papadopoulos et al. (2011), and $\gamma = 0, 0.1, 0.2, 0.3, \ldots, 2.0$,

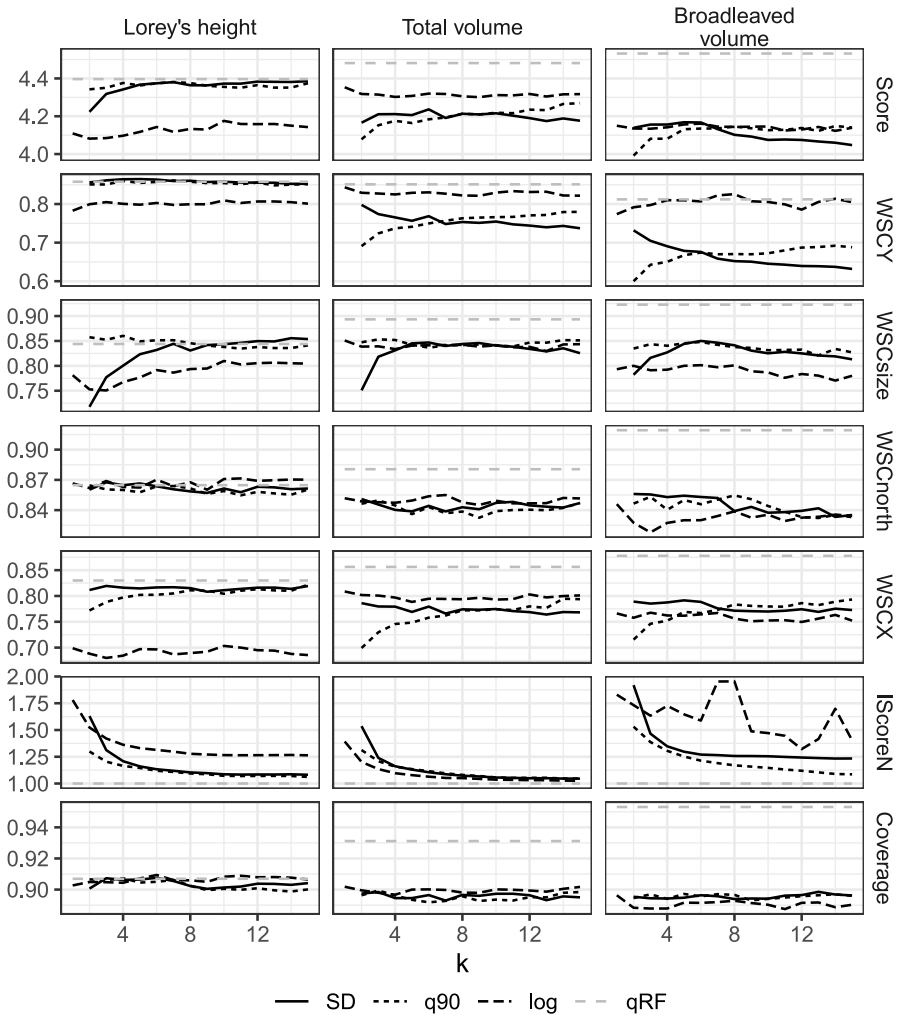

Fig. 5. Different evaluation metrics (rows) with respect to the number of neighbors k for the three forest attributes in Central Finland (columns) and four conformal procedures (lines; SD = Eq. (17) with $\gamma = 0.5$, q90 = quantile k-NN, log = log transform with optimized offset, qRF = quantile random forest). Please refer to Section 3.3 on the metric names.

- Conformal prediction with heuristic uncertainties based on the root mean squared error (Eq. (8)) of the k nearest neighbors, $u(x_i) = e(x_i)$, through the same transformations as $sd_k(x_i)$ and $d_k(x_i)$,
- Conformalized quantile regression with the lower and upper quantiles equal to the minimum and maximum (Eq. (11)) (q100), the $\alpha_{\text{lo}} = 0.05$ and $\alpha_{\text{hi}} = 0.95$ sample quantiles (q90), or the $\alpha_{\text{lo}} = 0.25$ and $\alpha_{\text{hi}} = 0.75$ sample quantiles (q50) from the k nearest neighbors.

The k-NN method as detailed in Section 3.1 was applied in all cases with different values of k from 2 up to 15. We compared the conformal k-NN predictions to the conformalized quantile regression using quantile random forest as implemented in the R package quantregForest (Meinshausen, 2006).

Both data sets (Finland, and Central Finland, see Fig. 3) were split into 70% training and 30% test data. For the split method, the training data was further split into the proper training set consisting of 60% of the full data and calibration set consisting of 10% of the full data. The k-NN based methods utilized the whole training data using jackknife, while the quantile random forest used split conformal prediction and trained the prediction model using the proper training set and calibrated the intervals using calibration set. Prediction intervals were predicted using the various methods listed above, and various efficiency measures from Section 3.3 were then computed from the test data. The train-test split was repeated ten times, and we report averages from these ten repeated runs.

3.5. Producing forest attribute maps with conformal k-NN prediction

To produce jackknife conformal k-NN predictions in a practical situation, one first needs to choose a conformal method to be used. If a method based on a heuristic uncertainty is chosen, e.g., the SD (Eq. (6)), the first step is to compute the predictions and heuristic uncertainties for each unit in the available training data. If a quantile conformal prediction is wished to be used, then the lower and upper quantile predictions for each unit are required. All the predictions should be made in the jackknife fashion, i.e., leaving the current unit i out from the training data. Having the required predictions, the required computations from the training data are the following:

1. Compute the conformity scores s_i for each unit i, $i = 1, \ldots, n$, from the training data.
2. Compute $\hat{q}$ from the s_i, $i = 1, \ldots, n$.

The quantile $\hat{q}$ will then be used to produce the prediction intervals for all grid cells $j = 1, \ldots, G$ in the study region. In the case of heuristic uncertainty, the two required steps are

1. For each grid cell j, $j = 1, \ldots, G$, compute the required heuristic uncertainty $u(x_j)$.
2. Compute the prediction intervals $C(x_j)$ by Eq. (5).

The equivalent steps for conformal quantile regression are:

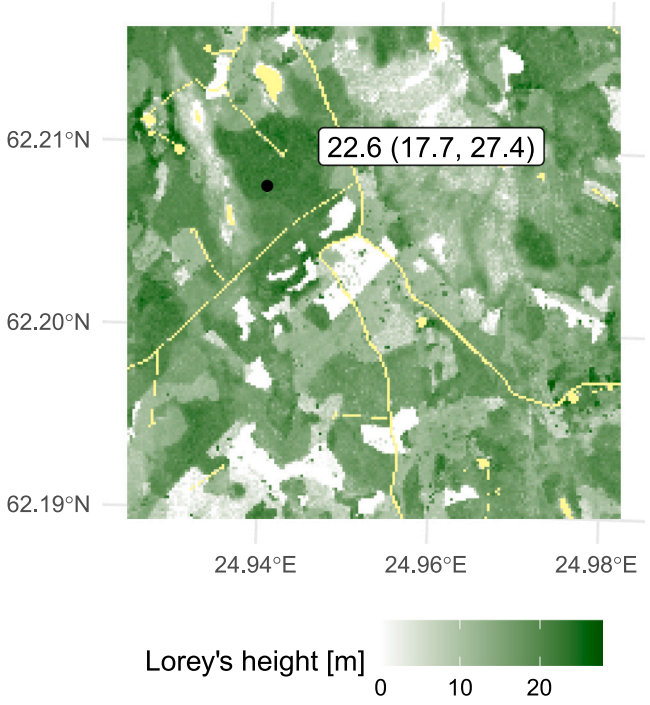

Fig. 6. Illustration of a potential online application for visualizing the map of Lorey's height based on the NFI field plots and ALS and satellite data from Central Finland shown for a subregion of size 3 km × 3 km. The method to produce the map was conformal k-NN with $k = 5$ and SD. Area outside forest is colored in yellow. Black dot denotes a selected grid cell and the text box displays the predicted Lorey's height and the associated 90% prediction interval.

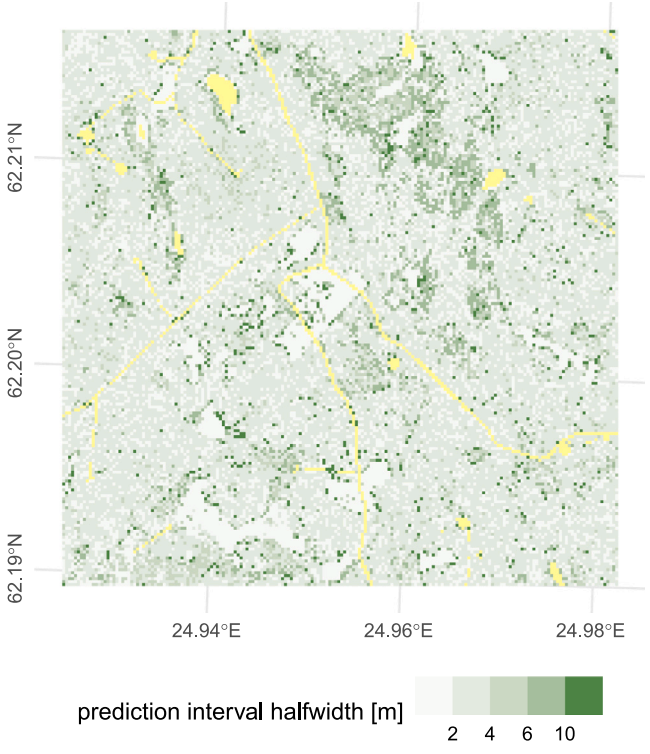

Fig. 7. The uncertainty layer of the map of Fig. 6. See Fig. 6 for an illustration of an online application.

1. For each grid cell j, $j = 1, \ldots, G$, compute the quantile predictions $\hat{q}_{\alpha_{\text{lo}}}(x_j)$ and $\hat{q}_{\alpha_{\text{hi}}}(x_j)$.
2. Compute the prediction intervals $C(x_j)$ by Eq. (10).

We note here that, if instead of the jackknife the split method is to be used, e.g., in the case of another prediction method, the training data are first split to $\mathcal{I}_1$ and $\mathcal{I}_2$ (see Section 3.2.1) and the s_i and $\hat{q}$ are computed only from $i \in \mathcal{I}_2$.

4. Results

4.1. The choice of tuning parameters

Some of the conformal k-NN prediction methods have tuning parameters. Namely, the logarithmic transformation (Eq. (16)) depends on λ, whereas the SD and mean distance based heuristic uncertainty measures depend on γ (see Section 3.4). These tuning parameters affect the performance of the conformal prediction intervals (Fig. B.8). The heuristic uncertainty measures $u(x_i)$ based on the SD or mean distance through an exponential transformation (Eq. (18)) were rather sensitive to the choice of the tuning parameter (Fig. B.8, middle), sometimes leading to absurdly long prediction intervals. On the other hand, the heuristic uncertainties with additive offsets behaved robustly with respect to the offset value (Fig. B.8, top). Thus, the transformations with additive offsets are to be preferred over the exponential transformations, and we will not consider the exponential transformations further; they also did not achieve higher overall scores than the transformations with additive offsets.

The overall score of the method based on the logarithmic transformation was also rather dependent on the offset (Fig. B.8, bottom). Due to the popularity of modeling the volume through logarithmic transformation, we anyway kept the logarithmic transformation in our further comparisons.

We further inspected the performance of the SD and additive offset (Eq. (17)) based method with respect to the different evaluation metrics (Fig. B.9). Generally the conditional coverage with respect to response (WSCY), northing (WSCnorth) or features (WSCX) was better for small offsets, and the interval length, conditional coverage and interval score were worse for small offsets. As a compromise we chose $\gamma = 0.5$ for the further comparisons, which was also used by Papadopoulos et al. (2011).

We did the same inspections for the logarithmic transformation, too (Fig. B.10). For the conformal procedure based on logarithmic transform, the choice of the offset parameter had a large effect on all evaluation metrics except WSCnorth. For total volume and Lorey's height all scores approximately agreed on which offset would result in the highest scores. The best offsets across the two study regions and different k according to the overall score were between 35 and 70 for total volume, and between 160 and 200 for Lorey's height. For broadleaves, especially the interval score was better for higher offset while other scores would choose small offsets between 0.7 and 4. For our further comparisons, we chose the optimal offset for each k, variable and region based on the overall score (15).

4.1.1. The choice of the k-NN quantiles

The conformalization allows using any quantiles for producing the desired (e.g., 90%) prediction intervals. We tested $(0.25, 0.75)$ (= q50), $(0.10, 0.90)$ (= q80), $(0.05, 0.95)$ (= q90) and the minimum and maximum (= q100) (Fig. B.11). Using q50 or q80 was clearly inferior to using q90 or q100. Using q100 led to slightly better evaluation metric values especially for larger k. However, this advantage is at least partially shadowed by a higher than 90% marginal coverage, i.e., on average overestimating the prediction uncertainty, for the broadleaved volume. Thus, we chose to use q90 in the following.

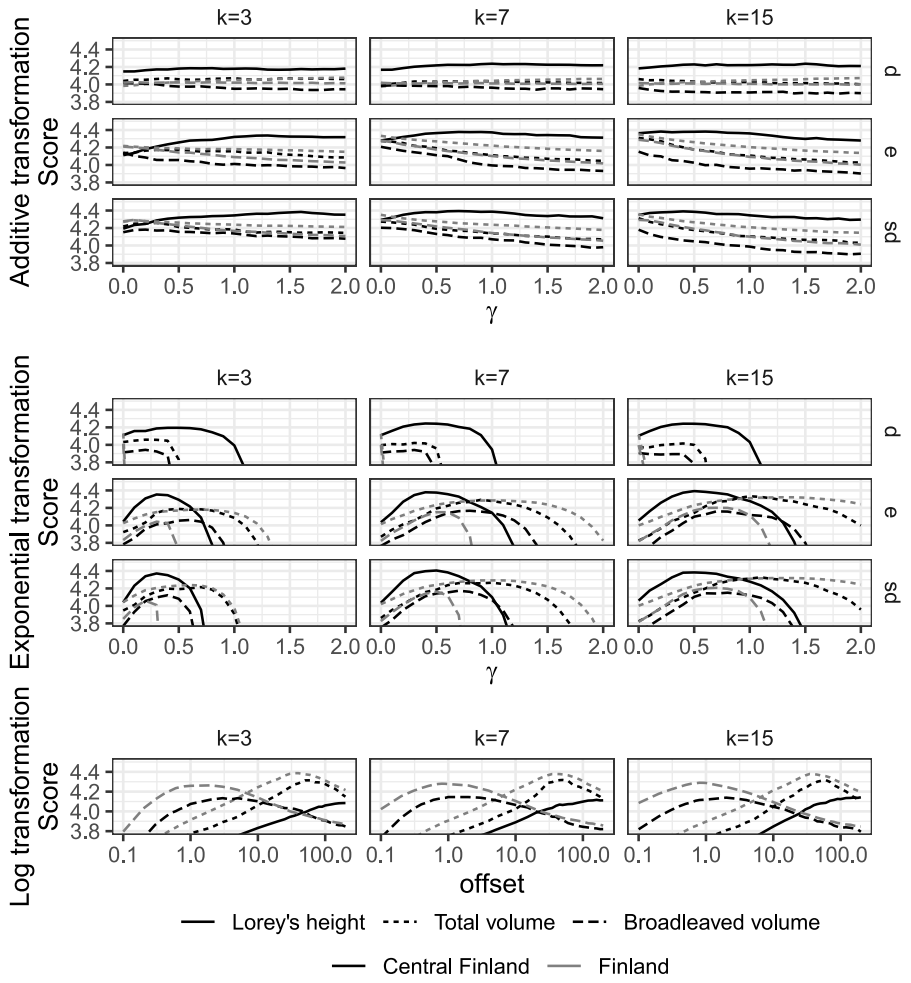

Fig. B.8. The overall score with respect to the tuning parameters of heuristic uncertainties using different transformations and statistics from the k nearest neighbors (d - Eq. (7); e - Eq. (8); sd - Eq. (6)). Columns give the results for different $k = 3, 7, 15$ and lines for the three forest attributes in two regions.

4.1.2. *Comparison of conformal prediction methods*

To our detailed comparisons across methods, we selected one heuristic uncertainty based conformal procedure (SD, additive transformation Eq. (17) with 0.5 offset), one quantile conformal procedure (q90) and one logarithmic transformation (log, optimized offset) based conformal procedure (see above). As a reference we used quantile random forest (qRF).

We computed the marginal coverage, different evaluation scores and the overall score (Eq. (15)) for all the methods both in Central Finland study region (Fig. 5) as well as the whole Finland (Fig. B.12). According to the overall score the log transform was a bit worse than SD and q90 for Lorey's height (top rows of Figs. 5 and Fig. B.12). On the other hand, the log transform was a bit better for total volume both in Finland and Central Finland and for the broadleaved volume in Finland. Regarding the behavior with respect to k, SD and q90 had a bit worse overall score for $k = 2$ than for $k \geq 3$: SD and quantiles cannot be well estimated from just two neighbors. Generally, q90 tended to get higher score for larger k, whereas SD started low at $k = 2$, reached a peak (dependent on the variable) and then the score started to decline as k grows. The overall score for the log transform was mostly independent of k.

For all variables and regions the reference procedure qRF achieved the best overall score. However, similarly as q100 (see Section 4.1.1 and Fig. B.11), qRF had overcoverage problems for broadleaved volume and to a small extent for total volume, i.e., the prediction intervals covered the ground truth more often than desired. This overcoverage partly explains higher evaluation metric values in these cases. All the compared k-NN procedures achieved marginal coverage that was very close to the target coverage 0.9 (bottom rows of Figs. 5 and B.12).

Inspecting the components of the overall score reveal further small differences between the methods: For the worst-slab coverage with respect to the response variable (WSCY) the log transform performed almost as good as qRF. The SD and q90 were worse for total volume and especially for broadleaved volume. Similarly as the overall score, the WSCY for q90 starts low and increases as k increases. The WSCY for SD peaks at $k = 2$ and then decreases as k grows. The WSCY for the log transform was mostly independent of k.

For the worst-slab coverage with respect to the interval length (WSCsize) q90 performed the best. The log transform performed as well as q90 for total volume, but it was the worst for the other variables. Here SD was the only procedure whose performance clearly varied with respect to k, having very low WSCsize for small values of k. That is, for small values of k, short and long intervals had different coverages.

The worst-slab coverage with respect to northing (WSCnorth) showed very little difference between the procedures.

The worst-slab coverage with respect to the model features (WSCX) behaved very much like WSCnorth. However, the log transform was worse than the other procedures for Lorey's height and better for total volume. Again q90 benefited from having a larger k.

For the interval score (IScoreN) the log transform was a bit worse for Lorey's height and broadleaved volume than SD and q90. On the other hand the log transform was better for total volume. q90 was always a bit better than SD. All procedures had worst interval scores for $k = 1$ and the score leveled out around $k = 5$. For the broadleaved volume and

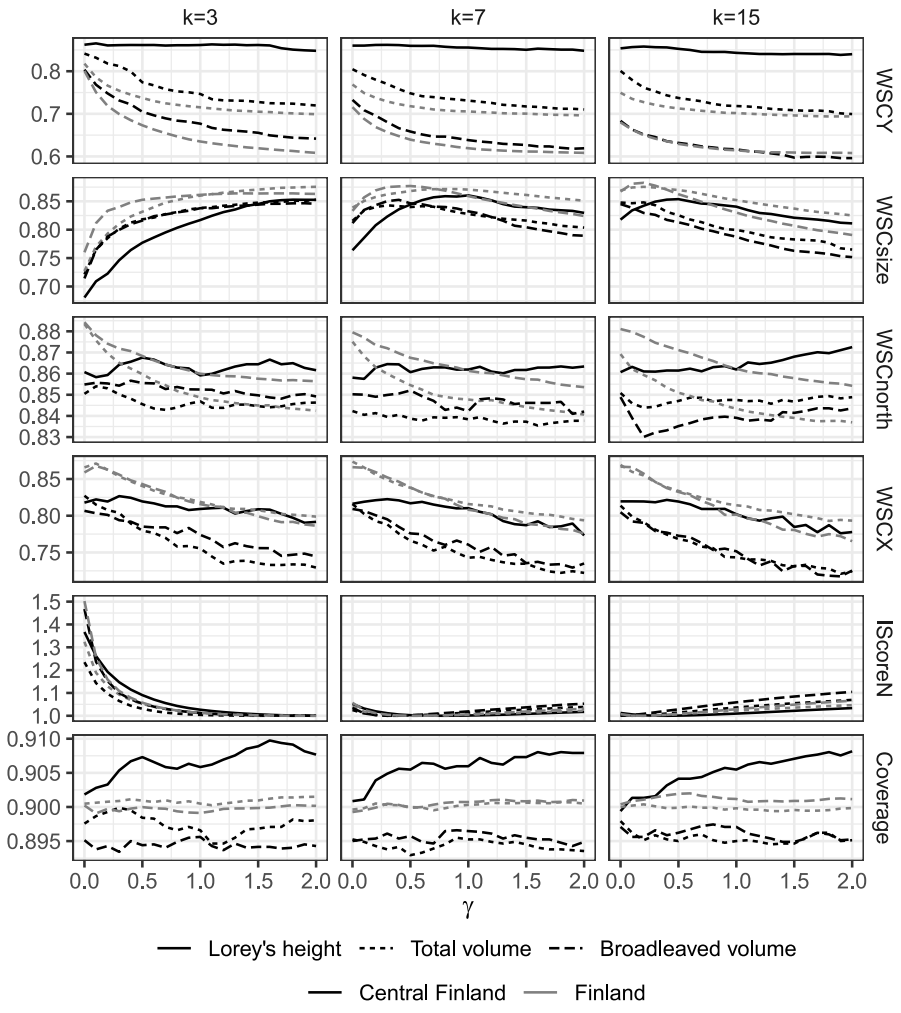

Fig. B.9. Different evaluation metrics (rows) with respect to the tuning parameter of the heuristic uncertainty based on SD (Eq. (17)). Columns give the results for different $k = 3, 7, 15$ and lines for the three forest attributes in two regions.

q90 the interval score kept getting better even up to $k = 15$. The interval score for log transform had some jumpy behavior for the broadleaved volume.

In conclusion, all conformal procedures performed rather well, but there were some minor differences between them with respect to some of the evaluation measures. Particularly, the smallest k produced larger interval scores. On the other hand, the performances with the different k varied with respect to the different worst-slab coverages.

4.2. Examples of conformal volume maps

We applied the k-NN method with $k = 5$ and using the SD based heuristic uncertainty measure Eq. (17) with $\gamma = 0.5$ to produce the predictions and uncertainties in the Central Finland. We followed the steps explained in Section 3.5. Fig. 6 shows the map of Lorey's height based on the NFI field plots and ALS and satellite data from Central Finland study area for a 3 km × 3 km subregion. The text box shows the predicted value with 90% prediction interval for the grid cell under the black dot. This illustrates how the uncertainties could be available for each grid cell in a potential online application. Fig. 7 shows the halfwidth of the prediction interval for the map in Fig. 6. The median halfwidth is 2.4 m. The prediction intervals tend to be wider on edges of forest stands or in locations with small-scale spatial variation with gaps in the canopy.

5. Discussion

Conformal prediction is a general technique to provide predictive sets that include the ground truth with at least a given probability. We proposed conformal predictive intervals to forest attribute mapping, as a way to provide grid cell level uncertainty quantification. We used satellite images and ALS data in our experiment but any remote sensing data suitable for mapping can be utilized. We focused on the k-NN method, which is popular in forest attribute mapping and for which no generally accepted method of uncertainty quantification is yet available. We proved a coverage guarantee for jackknife conformal k-NN prediction intervals (Appendix A) and showed by our experimental study (Section 4) that jackknife conformal prediction intervals using the k-NN method have good coverage properties. Particularly, the conformal predictions based on SD or quantiles of the k nearest neighbors had reasonable marginal and conditional coverages with commonly used values of k (as a compromise between RMSE and bias values $k < 10$ are commonly used, e.g., Mäkisara et al., 2022; Miettinen et al., 2024).

While conformal prediction method guarantees the desired marginal coverage, the heuristic uncertainty determines how well the prediction intervals adapt to the heteroscedasticity in data. In principle there are no restrictions for the source of heuristic uncertainty measures, but their usefulness certainly varies according to the ability of them to adapt to heteroscedasticity. In our experimental study, it turned out that some conformal procedures (e.g., using the logarithmic transformation) were rather sensitive to the tuning parameters. As we did not

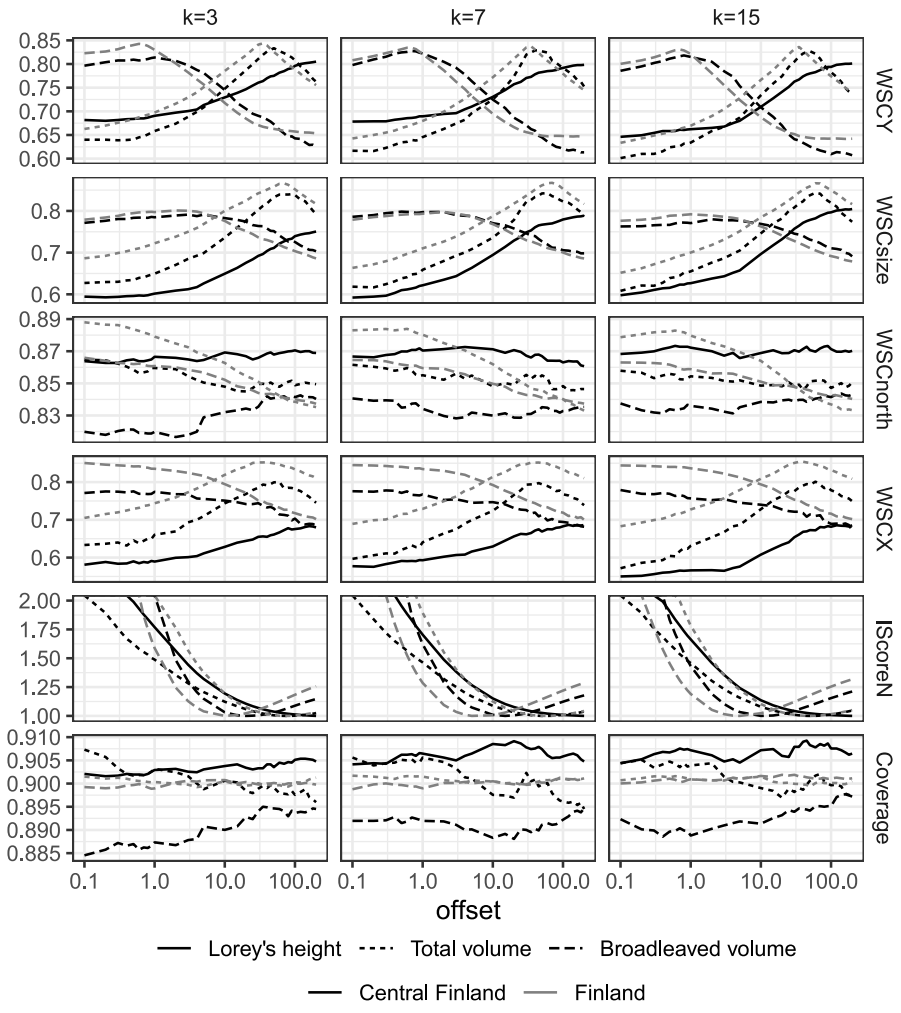

Fig. B.10. Different evaluation metrics (rows) with respect to the offset parameter of the logarithmic transformation for $k = 3, 7, 15$ (columns) and for the three forest attributes in two regions (lines).

find any obvious way to predict the best choice of the tuning parameters, we recommend using the procedures which were not sensitive to or did not have tuning parameters: the additive transformation of SD and the conformal quantile k-NN method.

We worked with the jackknife conformal prediction for the k-NN method. The jackknife method is attractive if representative training data are available and if the methods used for mapping allows effective use of it. We note that several common regression algorithms as well as RF allow, similarly as k-NN, to perform the jackknife conformal steps without the need to refit the model n times (Barber et al., 2021). While theoretical coverage guarantees for the jackknife method require some assumptions on the regression algorithm (Barber et al., 2021), the jackknife prediction intervals have been observed to achieve the target coverage in practice (Barber et al., 2021). Besides our detailed study with jackknife conformal k-NN prediction, we also tested the conformal jackknife method with qRF, using out-of-bag predictions for estimating the conformal adjustment ($\hat{q}$) and obtained results similar to those presented in Section 4 using qRF with split method. If the jackknife procedure is too computationally demanding, there exists a cross-validation based conformal prediction procedure and the split method. For stronger theoretical guarantees the split method is always available.

It should be noted that the calibration set used in conformal prediction requires careful consideration. The calibration set should be a representative sample of the target set. The simplest option for the calibration set is to have a simple random sample of the target population. In the context of forest attribute mapping, the training data are often systematic samples, stemming for example from NFIs. As the sample plots do not lie geographically too close to each other, they are typically considered to provide approximately independent observations, too. However, sample plots collected for a specific purpose may not serve the purpose.

When producing the map for a large region, there can be several sampling regions with different sampling designs. For example, the Finnish NFI in our study region (Fig. 3) consists of four sampling regions with different systematic designs. In our simulation study we worked only with the test data, and thus did not consider the sampling regions. However, the design can affect the validity of conformal prediction intervals in the map production. If the sampling intensities are not too different it may still be beneficial to apply conformal prediction, because the error made by not accounting for the sampling design is probably smaller than the error of using a heuristic uncertainty directly. Wieczorek et al. (2023) considered some complex sampling designs and proposed using so-called covariate shift method (Tibshirani et al., 2019). In future, we plan to investigate rigorous accounting for the sampling design when producing uncertainties for forest resource maps.

The assumption that the train and test data are identically distributed is in principle violated in the forest attribute mapping context due to several reasons: shape, size and time differences between field plots and remotely sensed data. Shape differences are typically believed to have a minor effect (Packalen et al., 2023). Here we worked with circular fields plots with radius 9 m and grid cells of size 16 m ×

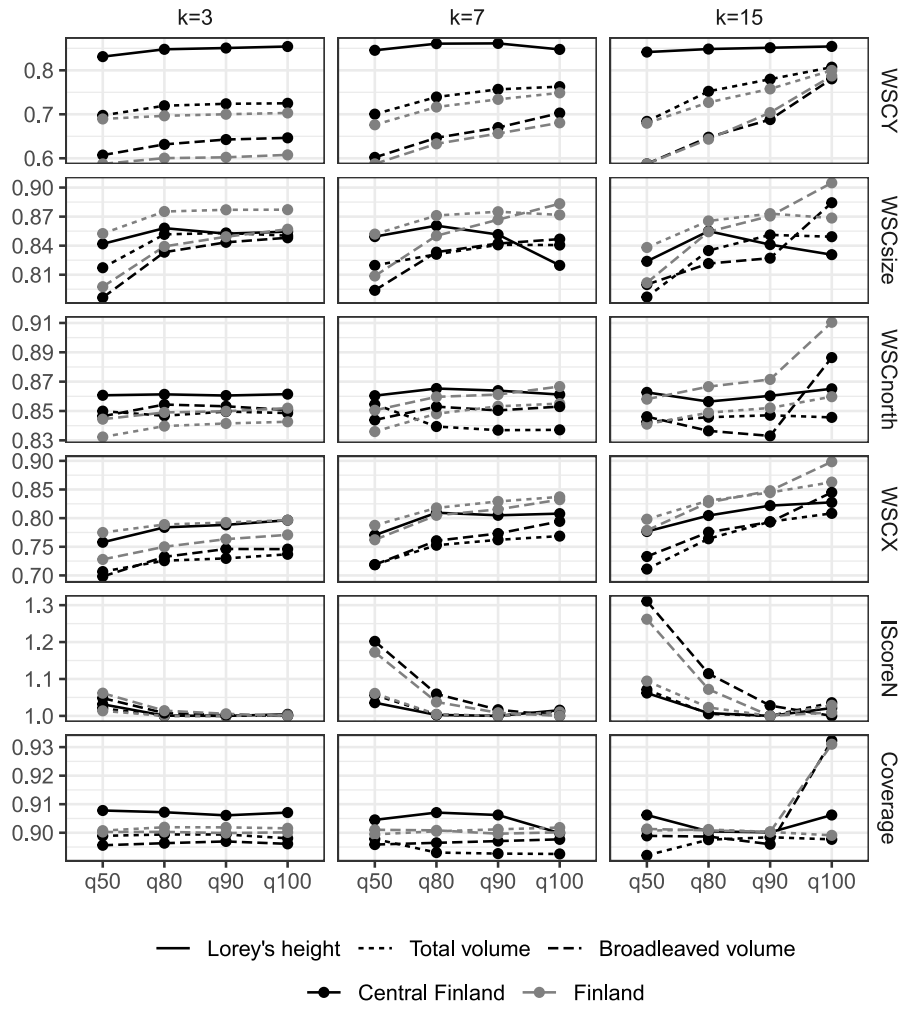

Fig. B.11. Different evaluation metrics (rows) for the conformal k-NN quantile methods for $k = 3, 7, 15$ (columns) and for the three forest attributes in two regions (lines).

16 m, which are approximately of the same size. Much greater size differences may play some role. The issue with time differences is that some cuttings can occur between the field data collection and remote sensing acquisition. On the other hand, the prediction map is made for the exact dates of the remote sensing features. This introduces some difference in the distribution of the training set and the prediction set. To mitigate the time issue, we removed field data where a clear cut was detected based on the field measured volume and the height from remote sensing data. Future work could investigate the possibility to include the time difference (δT) in a heuristic uncertainty measure, for example $u(x) = sd(x) + \delta T + \gamma$.

According to a typical coverage result, the coverage is at least $1 - \alpha$. There is an additional result that the coverage is less than $1 - \alpha + 1/n$ which requires continuous distribution of the errors/scores. When there are a lot of zero volumes the continuity is not satisfied. For the broadleaved volume that is quite often zero, we observed overcoverage of the conformal prediction intervals for some conformal quantile prediction procedures, namely the conformal quantile random forest (the reference method) and the conformal k-NN quantiles based on the minimum and maximum value of the k nearest neighbors (q100). The overcoverage is a result of perfectly predicting the lower bound of the prediction interval at zero for majority of the zeros in the data, thus leaving little space for adjusting the coverage to the desired level. Future work should investigate scores specialized to the case of having lots of zeros, since zeros are quite common with forest attributes.

In conclusion, there are several things that can still enhance the conformal predictions. Still we believe that already the basic approach used here is beneficial in many situations.

Finally, similar to a typical case in forest resource mapping, our field sample plots were relatively small and did not lie in the close vicinity of each other. Thus, we were able to conformalize our predictions at the unit level, i.e., in the areas of the same size as the field plot data which we had available for training. In the future, it would be interesting to investigate whether or not conformal prediction could help also in small-area estimation context. This might however require training data that are currently not available.

CRediT authorship contribution statement

M. Kuronen: Writing – review & editing, Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **J. Räty:** Writing – review & editing, Visualization, Methodology, Investigation, Formal analysis, Data curation. **P. Packalen:** Writing – review & editing, Methodology, Investigation. **M. Myllymäki:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Investigation, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The research leading to these results has received funding from the European Union Horizon Europe (HORIZON) Research & Innovation programme under the Grant Agreement no. 101056907 (PathFinder), and from the European Union—NextGenerationEU in the Research Council of Finland's project (Grant number 348154). The work was done under the Research Council of Finland's flagship ecosystem for Forest-Human-Machine Interplay—Building Resilience, Redefining Value Networks and Enabling Meaningful Experiences (UNITE) (Grant number 357909). We thank our colleagues Timo P. Pitkänen for the Sentinel-2 mosaic, and Juho Pitkänen and Andras Balazs for help on data handling.

Appendix A. Proof of Eq. (12)

We say that a score function is *out-of-sample stable* for $\nu \in [0, 1]$ if

$$P(\hat{s}(X_{n+1}, Y_{n+1}) \neq \hat{s}_{-i}(X_{n+1}, Y_{n+1})) < \nu, \tag{19}$$

i.e., the predictions for a new data point (X_{n+1}, Y_{n+1}) with using also the ith observation, or not using it, are the same with a high probability. This condition is analogous to the out-of-sample stability condition in Barber et al. (2021).

Suppose that the score function satisfies the out-of-sample stability condition (Eq. (19)) with ν. Below we will show that then the jackknife prediction interval satisfies

$$P(Y_{n+1} \in \hat{C}(X_{n+1})) \geq 1 - \alpha - 2\sqrt{\nu}. \tag{20}$$

A conformity score that is based on k nearest neighbors is out-of-sample stable for $\nu = k/n$. This can be easily seen by using the arguments in Barber et al. (2021, Sec. 5.5). Thus, in summary, the coverage of the prediction intervals is at least $1 - \alpha - 2\sqrt{k/n}$.

The proof closely follows the proof of Theorem 5 in Barber et al. (2021). We start by showing that jackknife using oracle scores achieves the desired coverage rate: For each $i = 1, \ldots, n+1$, let $\tilde{s}_{-i}$ be the oracle score function based on the training and the new data point with point i removed, i.e.,

$$\tilde{s}_{-i} = \mathcal{A}((X_1, Y_1), \ldots, (X_{i-1}, Y_{i-1}), (X_{i+1}, Y_{i+1}), \ldots, (X_{n+1}, Y_{n+1})).$$

These functions are used to compute $\tilde{q}$, the $\lceil (1-\alpha')(n+1) \rceil / n$ quantile of the scores $\tilde{s}_{-i}(x_i, y_i)$, $i = 1, \ldots, n$, where $\alpha' = \alpha + \sqrt{\nu}$. The oracle prediction interval is now

$$\tilde{C}(x) = \{y : \hat{s}(x, y) \leq \tilde{q}\}.$$

This interval by definition satisfies

$$P(Y_{n+1} \in \tilde{C}(X_{n+1})) = P(\hat{s}(Y_{n+1}, X_{n+1}) \leq \tilde{q}).$$

Now since the training and new data point are assumed to be i.i.d. and the function $\mathcal{A}$ to be symmetric on its arguments, the oracle scores

$$\tilde{s}_{-1}(X_1, Y_1), \ldots, \tilde{s}_{-(n+1)}(X_{n+1}, Y_{n+1})$$

are also exchangeable. Thus, because $\tilde{s}_{-(n+1)}(X_{n+1}, Y_{n+1}) = \hat{s}(X_{n+1}, Y_{n+1})$,

$$P(\hat{s}(X_{n+1}, Y_{n+1}) \leq \tilde{q}) \geq \frac{\lceil (1-\alpha')(n+1) \rceil}{n+1} \geq 1 - \alpha'.$$

In the second part we show that the oracle interval is with sufficient probability covered by the jackknife interval if the score function is out-of-sample stable. The inclusion holds if $\tilde{q} \leq \hat{q}$. If $\tilde{q} > \hat{q}$ then the number of indices $i = 1, \ldots, n$ with $\tilde{s}_{-i} \neq \hat{s}_{-i}$ is at least

$$\lceil (1-\alpha)(n+1) \rceil - \lceil (1-\alpha')(n+1) \rceil + 1 \geq \sqrt{\nu}(n+1).$$

Therefore

$$\begin{aligned} P(\tilde{C}(X_{n+1}) \not\subset \hat{C}(X_{n+1})) &= P(\tilde{q} > \hat{q}) \\ &\leq P\Big(\sum_{i=1}^{n} 1(\tilde{s}_{-i} \neq \hat{s}_{-i}) \geq \sqrt{\nu}(n+1)\Big) \\ &\leq \frac{E(\sum_{i=1}^{n} 1(\tilde{s}_{-i} \neq \hat{s}_{-i}))}{\sqrt{\nu}(n+1)}, \end{aligned}$$

where the last inequality holds by Markov's inequality. For every $i = 1, \ldots, n$

$$\begin{aligned} P(\tilde{s}_{-i}(X_i, Y_i) \neq \hat{s}_{-i}(X_i, Y_i)) &= P(\tilde{s}_{-i}(X_i, Y_i) \neq \tilde{s}_{-(i,n+1)}(X_i, Y_i)) \\ &= P(\tilde{s}_{-(n+1)}(X_{n+1}, Y_{n+1}) \neq \tilde{s}_{-(n+1,i)}(X_{n+1}, Y_{n+1})) \\ &= P(\hat{s}(X_{n+1}, Y_{n+1}) \neq \hat{s}_{-i}(X_{n+1}, Y_{n+1})) \\ &\leq \nu, \end{aligned}$$

where the first and third step hold by the definition of the score functions (in $\tilde{s}_{-(i,n+1)}(X_i, Y_i)$ both ith and $(n+1)$th point is excluded), the second step holds since the data points are iid, and the last step holds due to out-of-sample stability. Thus

$$P(\tilde{C}(X_{n+1}) \not\subset \hat{C}(X_{n+1})) \leq \frac{\nu n}{\sqrt{\nu}(n+1)} \leq \sqrt{\nu}$$

Combining the two parts gives

$$\begin{aligned} P(Y_{n+1} \in \hat{C}(X_{n+1})) &\geq P(Y_{n+1} \in \tilde{C}(X_{n+1})) - P(\tilde{C}(X_{n+1}) \not\subset \hat{C}(X_{n+1})) \\ &\geq 1 - \alpha - 2\sqrt{\nu}. \end{aligned}$$

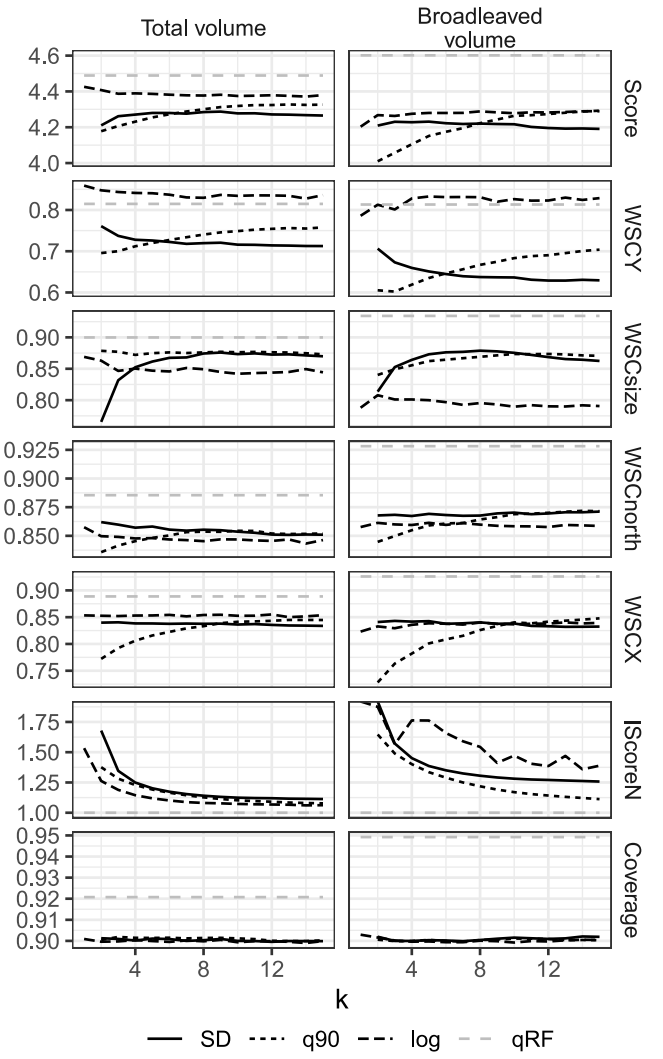

Fig. B.12. Different evaluation metrics (rows) with respect to the number of neighbors k for the two forest attributes in Finland (columns) and four conformal procedures (lines).

Appendix B. Additional results for the simulation study

See Figs. B.8–B.12.

Data availability

The authors do not have permission to share data.

References

Angelopoulos, A.N., Bates, S., 2022. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. http://dx.doi.org/10.48550/arXiv.2107.07511, arXiv:2107.07511 [cs, math, stat].

Axelsson, P., 2000. DEM generation from laser scanner data using adaptive TIN models. Int. Arch. Photogramm. Remote Sens. 33 (4), 110–117.

Barber, R.F., Candès, E.J., Ramdas, A., Tibshirani, R.J., 2021. Predictive inference with the jackknife+. Ann. Statist. 49 (1), http://dx.doi.org/10.1214/20-AOS1965.

Baston, D., 2023. exactextractr: Fast extraction from raster datasets using polygons. URL https://CRAN.R-project.org/package=exactextractr.

Breidenbach, J., Ellison, D., Petersson, H., Korhonen, K.T., Henttonen, H.M., Wallerman, J., Fridman, J., Gobakken, T., Astrup, R., Næsset, E., 2022. Harvested area did not increase abruptly—how advancements in satellite-based mapping led to erroneous conclusions. Ann. For. Sci. 79 (1), 2. http://dx.doi.org/10.1186/s13595-022-01120-4.

Cauchois, M., Gupta, S., Duchi, J.C., 2021. Knowing what you know: valid and validated confidence sets in multiclass and multilabel prediction. J. Mach. Learn. Res. 22 (1), 81:3681–81:3722.

Chirici, G., Mura, M., McInerney, D., Py, N., Tomppo, E.O., Waser, L.T., Travaglini, D., McRoberts, R.E., 2016. A meta-analysis and review of the literature on the k-Nearest Neighbors technique for forestry applications that use remotely sensed data. Remote Sens. Environ. 176, 282–294. http://dx.doi.org/10.1016/j.rse.2016.02.001.

Cosenza, D.N., Korhonen, L., Maltamo, M., Packalen, P., Strunk, J.L., Næsset, E., Gobakken, T., Soares, P., Tomé, M., 2021. Comparison of linear regression, k-nearest neighbour and random forest methods in airborne laser-scanning-based prediction of growing stock. For.: An Int. J. For. Res. 94 (2), 311–323. http://dx.doi.org/10.1093/forestry/cpaa034.

Crookston, N.L., Finley, A.O., 2007. yaImpute: An R package for kNN imputation. J. Stat. Softw. 23 (10), http://dx.doi.org/10.18637/jss.v023.i10.

Fassnacht, F.E., White, J.C., Wulder, M.A., Næsset, E., 2024. Remote sensing in forestry: current challenges, considerations and directions. For.: An Int. J. For. Res. 11–37. http://dx.doi.org/10.1093/forestry/cpad024.

Gneiting, T., Raftery, A.E., 2007. Strictly proper scoring rules, prediction, and estimation. J. Amer. Statist. Assoc. 102 (477), 359–378. http://dx.doi.org/10.1198/016214506000001437.

Hyndman, R.J., Fan, Y., 1996. Sample quantiles in statistical packages. Amer. Statist. 50 (4), 361–365. http://dx.doi.org/10.2307/2684934.

Kakhani, N., Alamdar, S., Kebonye, N.M., Amani, M., Scholten, T., 2024. Uncertainty quantification of soil organic carbon estimation from remote sensing data with conformal prediction. Remote. Sens. 16 (3), 438. http://dx.doi.org/10.3390/rs16030438.

Kangas, A., Astrup, R., Breidenbach, J., Fridman, J., Gobakken, T., Korhonen, K.T., Maltamo, M., Nilsson, M., Nord-Larsen, T., Næsset, E., Olsson, H., 2018. Remote sensing and forest inventories in nordic countries – roadmap for the future. Scand. J. For. Res. 33 (4), 397–412. http://dx.doi.org/10.1080/02827581.2017.1416666.

Kangas, A., Myllymäki, M., Mehtätalo, L., 2023. Understanding uncertainty in forest resources maps. Silva Fenn. 57 (2), http://dx.doi.org/10.14214/sf.22026.

Kangas, A., Myllymäki, M., Packalen, P., 2024. Small area estimators in a simulation test. Can. J. Forest Res. http://dx.doi.org/10.1139/cjfr-2024-0070, Publisher: NRC Research Press.

Kim, H.-J., Tomppo, E., 2006. Model-based prediction error uncertainty estimation for k-nn method. Remote Sens. Environ. 104 (3), 257–263. http://dx.doi.org/10.1016/j.rse.2006.04.009.

Kirkpatrick, S., Gelatt, C.D., Vecchi, M.P., 1983. Optimization by simulated annealing. Science 220 (4598), 671–680. http://dx.doi.org/10.1126/science.220.4598.671.

Korhonen, K.T., Ahola, A., Heikkinen, J., Henttonen, H.M., Hotanen, J.-P., Ihalainen, A., Melin, M., Pitkänen, J., Räty, M., Sirviö, M., Strandström, M., 2021. Forests of Finland 2014–2018 and their development 1921–2018. Silva Fenn. 55 (5), http://dx.doi.org/10.14214/sf.10662.

Korhonen, K.T., Räty, M., Haakana, H., Heikkinen, J., Hotanen, J.-P., Kuronen, M., Pitkänen, J., 2024. Forests of Finland 2019–2023 and their development 1921–2023. Silva Fenn. 58 (5), http://dx.doi.org/10.14214/sf.24045.

Kotivuori, E., Maltamo, M., Korhonen, L., Strunk, J.L., Packalen, P., 2021. Prediction error aggregation behaviour for remote sensing augmented forest inventory approaches. For.: An Int. J. For. Res. 94 (4), 576–587. http://dx.doi.org/10.1093/forestry/cpab007.

Laasasenaho, J., 1982. Taper Curve and Volume Functions for Pine, Spruce and Birch, vol. 108, Communicationes Instituti Forestalia Fennica, URL http://urn.fi/URN:ISBN:951-40-0589-9.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R.J., Wasserman, L., 2018. Distribution-free predictive inference for regression. J. Amer. Statist. Assoc. 113 (523), 1094–1111. http://dx.doi.org/10.1080/01621459.2017.1307116.

Mäkisara, K., Katila, M., Peräsaari, J., 2022. The Multi-Source National Forest Inventory of Finland — Methods and Results 2017 and 2019. Natural Resources Institute Finland (Luke), URL http://urn.fi/URN:ISBN:978-952-380-538-5.

Maltamo, M., Packalen, P., 2014. Species-specific management inventory in Finland. In: Maltamo, M., Næsset, E., Vauhkonen, J. (Eds.), Forestry Applications of Airborne Laser Scanning: Concepts and Case Studies. Springer Netherlands, Dordrecht, pp. 241–252. http://dx.doi.org/10.1007/978-94-017-8663-8_12.

McRoberts, R.E., Magnussen, S., Tomppo, E.O., Chirici, G., 2011. Parametric, bootstrap, and jackknife variance estimators for the k-Nearest Neighbors technique with illustrations using forest inventory and satellite image data. Remote Sens. Environ. 115 (12), 3165–3174. http://dx.doi.org/10.1016/j.rse.2011.07.002.

McRoberts, R.E., Tomppo, E.O., 2007. Remote sensing support for national forest inventories. Remote Sens. Environ. 110 (4), 412–419. http://dx.doi.org/10.1016/j.rse.2006.09.034.

McRoberts, R.E., Tomppo, E.O., Næsset, E., 2010. Advances and emerging issues in national forest inventories. Scand. J. For. Res. 25 (4), 368–381. http://dx.doi.org/10.1080/02827581.2010.496739.

Meinshausen, N., 2006. Quantile regression forests. J. Mach. Learn. Res. 7, 983–999.

Meyer, H., Pebesma, E., 2021. Predicting into unknown space? Estimating the area of applicability of spatial prediction models. Methods Ecol. Evol. 12 (9), 1620–1633. http://dx.doi.org/10.1111/2041-210X.13650.

Meyer, H., Pebesma, E., 2022. Machine learning-based global maps of ecological variables and the challenge of assessing them. Nat. Commun. 13 (1), 2208. http://dx.doi.org/10.1038/s41467-022-29838-9.

Miettinen, J., Adame, P., Adolt, R., Alberdi, I., Antropov, O., Arnarsson, Ó., Astrup, R., Berger, A., Bogason, J., Chirici, G., Corona, P., D'Amico, G., Fejfar, J., Fischer, C., Gohon, F., Gschwantner, T., Hertzler, J., Koma, Z., Korhonen, K.T., Krajnc, L., Latte, N., Lejeune, P., McCullagh, A., Mionskowski, M., Moreno, D., Myllymäki, M., Nilsson, M., Perin, J., Pitkänen, J., Redmond, J., Riedel, T., Schumacher, J., Seitsonen, L., Sirro, L., Skudnik, M., Snorrason, A., Sroga, R.a., Traub, B., Westerlund, B., Wurpillot, S., Breidenbach, J., 2024. High-resolution pan-European forest structure maps: An integration of earth observation and national forest inventory data. http://dx.doi.org/10.5281/zenodo.13143235, URL https://zenodo.org/records/13143235.

Mitchard, E.T., Saatchi, S.S., Baccini, A., Asner, G.P., Goetz, S.J., Harris, N.L., Brown, S., 2013. Uncertainty in the spatial distribution of tropical forest biomass: a comparison of pan-tropical maps. Carbon Balance Manag. 8 (1), 10. http://dx.doi.org/10.1186/1750-0680-8-10.

National Land Survey of Finland, 2024. Laser scanning data 0,5p. Retrieved November 25, 2024, from https://www.maanmittauslaitos.fi/en/maps-and-spatial-data/datasets-and-interfaces/product-descriptions/laser-scanning-data-05-p.

Norinder, U., Lowry, S., 2023. Predicting larch casebearer damage with confidence using Yolo network models and conformal prediction. Remote. Sens. Lett. 14 (10), 1021–1033. http://dx.doi.org/10.1080/2150704X.2023.2258460.

Packalen, P., Strunk, J., Maltamo, M., Myllymäki, M., 2023. Circular or square plots in ALS-based forest inventories—does it matter? For.: An Int. J. For. Res. 96 (1), 49–61. http://dx.doi.org/10.1093/forestry/cpac032.

Packalén, P., Temesgen, H., Maltamo, M., 2012. Variable selection strategies for nearest neighbor imputation methods used in remote sensing based forest inventory. Can. J. Remote Sens. 38 (5), 557–569. http://dx.doi.org/10.5589/m12-046.

Palahí, M., Valbuena, R., Senf, C., Acil, N., Pugh, T.A.M., Sadler, J., Seidl, R., Potapov, P., Gardiner, B., Hetemäki, L., Chirici, G., Francini, S., Hlásny, T., Lerink, B.J.W., Olsson, H., González Olabarria, J.R., Ascoli, D., Asikainen, A., Bauhus, J., Berndes, G., Donis, J., Fridman, J., Hanewinkel, M., Jactel, H., Lindner, M., Marchetti, M., Marušák, R., Sheil, D., Tomé, M., Trasobares, A., Verkerk, P.J., Korhonen, M., Nabuurs, G.-J., 2021. Concerns about reported harvests in European forests. Nature 592 (7856), E15–E17. http://dx.doi.org/10.1038/s41586-021-03292-x.

Papadopoulos, H., Gammerman, A., Vovk, V., 2008. Normalized nonconformity measures for regression conformal prediction. In: Proceedings of the 26th IASTED International Conference on Artificial Intelligence and Applications. AIA '08, ACTA Press, USA, pp. 64–69.

Papadopoulos, H., Vovk, V., Gammerman, A., 2011. Regression conformal prediction with nearest neighbours. J. Artificial Intelligence Res. 40, 815–840. http://dx.doi.org/10.1613/jair.3198.

Pitkänen, T.P., Balazs, A., Tuominen, S., 2024. Automatized sentinel-2 mosaicking for large area forest mapping. Int. J. Appl. Earth Obs. Geoinf. 127, 103659. http://dx.doi.org/10.1016/j.jag.2024.103659.

R Core Team, 2023. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, URL https://www.R-project.org/.

Romano, Y., Patterson, E., Candes, E., 2019. Conformalized quantile regression. Adv. Neural Inf. Process. Syst.

Sagar, A., Vega, C., Bouriaud, O., Piedallu, C., Renaud, J.-P., 2022. Multisource forest inventories: A model-based approach using *k*-NN to reconcile forest attributes statistics and map products. ISPRS J. Photogramm. Remote Sens. 192, 175–188. http://dx.doi.org/10.1016/j.isprsjprs.2022.08.016.

Schulp, C.J.E., Burkhard, B., Maes, J., Vliet, J.V., Verburg, P.H., 2014. Uncertainties in ecosystem service maps: A comparison on the European scale. Plos One 9 (10), e109643. http://dx.doi.org/10.1371/journal.pone.0109643.

Singh, G., Moncrieff, G., Venter, Z., Cawse-Nicholson, K., Slingsby, J., Robinson, T.B., 2024. Uncertainty quantification for probabilistic machine learning in earth observation using conformal prediction. Sci. Rep. 14 (1), 16166. http://dx.doi.org/10.1038/s41598-024-65954-w.

Tibshirani, R.J., Barber, R.F., Candès, E.J., Ramdas, A., 2019. Conformal prediction under covariate shift. Adv. Neural Inf. Process. Syst. 32.

Tomppo, E., Halme, M., 2004. Using coarse scale forest variables as ancillary information and weighting of variables in k-NN estimation: a genetic algorithm approach. Remote Sens. Environ. 92 (1), 1–20. http://dx.doi.org/10.1016/j.rse.2004.04.003.

Valle, D., Izbicki, R., Leite, R.V., 2023. Quantifying uncertainty in land-use land-cover classification using conformal statistics. Remote Sens. Environ. 295, 113682. http://dx.doi.org/10.1016/j.rse.2023.113682.

Vovk, V., 2013. Conditional validity of inductive conformal predictors. Mach. Learn. 92 (2–3), 349–376. http://dx.doi.org/10.1007/s10994-013-5355-6.

Vovk, V., 2015. Cross-conformal predictors. Ann. Math. Artif. Intell. 74 (1), 9–28. http://dx.doi.org/10.1007/s10472-013-9368-4.

Wieczorek, J.A., White, G.W., Cody, Z.W., Tan, E.X., Chistolini, J.O., McConville, K.S., Frescino, T.S., Moisen, G.G., 2023. Assessing small area estimates via artificial populations from KBAABB: a kNN-based approximation to ABB. http://dx.doi.org/10.48550/arXiv.2306.15607, arXiv:2306.15607 [stat].

Zhang, Y., Liang, S., Yang, L., 2019. A review of regional and global gridded forest biomass datasets. Remote. Sens. 11 (23), 2744. http://dx.doi.org/10.3390/rs11232744.