forests

MDPI

Article

Bayesian Approach for Optimizing Forest Inventory Survey Sampling with Remote Sensing Data

Jonne Pohjankukka [1,*], Sakari Tuominen [2] and Jukka Heikkonen [1]

[1] Department of Computing, University of Turku, Vesilinnantie 5, FI-20500 Turku, Finland
[2] Natural Resources Institute Finland (LUKE), Latokartanonkaari 9, FI-00790 Helsinki, Finland
* Correspondence: jjepoh@utu.fi

Abstract: In large-area forest inventories, a trade-off between the amount of data to be sampled and the corresponding collection costs is necessary. It is not always possible to have a very large data sample when dealing with sampling-based inventories. It is therefore important to optimize the sampling design with the limited resources. Whereas this sort of inventories are subject to these constraints, the availability of remote sensing (RS) data correlated with the forest inventory variables is usually much higher. For this reason, the RS and sampled field measurement data are often used in combination for improving the forest inventory estimation. In this study, we propose a model-based data sampling method founded on Bayesian optimization and machine learning algorithms which utilizes RS data to guide forest inventory sample selection. We evaluate our method in empirical experiments using real-world volume of growing stock data from the Aland region in Finland. The proposed method is compared against two baseline methods: simple random sampling and the local pivotal method. When a suitable model link is selected, the empirical experiments show on best case on average up to 22% and 79% improvement in population mean and variance estimation respectively over baselines. However, the results also illustrate the importance of model selection which has a clear effect on the results. The novelty of the study is in the application of Bayesian optimization in national forest inventory survey sampling.

Keywords: national forest inventory; remote sensing; survey sampling; Bayesian optimization; machine learning

Citation: Pohjankukka, J.; Tuominen, S.; Heikkonen, J. Bayesian Approach for Optimizing Forest Inventory Survey Sampling with Remote Sensing Data. *Forests* **2022**, *13*, 1692. https://doi.org/10.3390/f13101692

Academic Editor: Gherardo Chirici

Received: 14 September 2022
Accepted: 4 October 2022
Published: 14 October 2022

1. Introduction

Large-area surveys apply a wide range of methods from the field of statistical sampling theory, such as, for example, the simple random, systematic or stratified sampling [1–6]. Methods such as systematic or clustered sampling are common in forest inventories, because the weights of individual sample plots are constant, which makes their application straightforward in monitoring the forest resources over consecutive inventories. The sampling methodologies are optimized in order to produce accurate inventories for the response variables of interest. Large-area forest inventories at regional and national level are typically based on sampled field observations measured from sample plots. The sampling intensity is dependent on the size of the inventory area, desired accuracy of the inventory data and the resources available for measuring the data. The sampled data should be representative enough to cover the variation of the significant variables, such as the volume of growing stock and main tree species in the inventory area, in order to allow the estimation of these variables at national and regional level. Management of forest resources requires predictions, e.g., on the distribution of tree species, state of forests, soil conditions for trafficability assessment etc. [7–10]. Information gain from these data-based approaches will be utilized both in strategic and operative plannings in forestry. Increasing the number of field observations generally improves the precision and accuracy of the inventory data, but on the other hand, the measurement of field data is the most significant cost factor in

sampling-based forest inventories. Thus, the selected inventory design is always a trade-off between the desired accuracy of inventory data and the available resources. One output of forest inventories, with growing importance, are thematic maps. These thematic maps are created using the collected field sample data and some prediction model, e.g., k-nearest neighbor, for interpolating the field sample data over larger areas [11].

The efficiency of sampling designs can be improved by using auxiliary data such as remote sensing (RS) data, which as such is not accurate enough for the inventory task but which can be used for enhancing the sampling efficiency by, e.g., weighting the areas represented by each sample plot. The main prerequisite for the use of auxiliary data is that there is sufficient correlation between the auxiliary data and actual variables of interest, which typically is the case between RS data and forest inventory variables [12–19]. Recent examples in Swedish forest inventory utilizing auxiliary information in sampling decisions have been presented, e.g., in the studies by [20–23]. In the works of [24,25] auxiliary data were used via the local pivotal method in national forest inventory (NFI) using Southern Finland as the test area. The local pivotal method produces sample locations in a stochastic manner while trying to avoid similar data points in the auxiliary space to be included into the data set, in order to produce a spatially well-balanced data set. The results showed significant improvement in estimation accuracy with the utilization of auxiliary data to the NFI. In addition, related studies regarding the baseline methods used in this work and sampling strategies regarding the NFI can be found, e.g., in the works by [26–28].

In design-based inference, sample locations are decided using probability sampling which means that each sample point (i.e., a single observation of a variable of interest in a spatial location) is included into the sample randomly based on inclusion probabilities, and the specific probability sampling method applied. In model-based inference, sample locations are not based on probability sampling but rather on a statistical model. Both of these approaches have their strengths and weaknesses, and the best approach of these two depends on the aim of the survey [29]. Primarily, the design-based approach is the most appropriate if interest is in the valid objective assessment of the population mean, whereas the model-based approach is the more appropriate if the goal is to map the study/response variable. In addition, given that a good model is used, the model-based approach is also more efficient with more precise estimates of the population mean [30] (ch. 26). In this work, we use a model-based sampling approach in which we utilize the model link between the response and auxiliary data for deciding new field sample locations. Our response variables of interest is the volume of growing stock, and the auxiliary data consist of remotely sensed airborne imagery and laser scanning data. The goal of the sampling is to estimate the population parameters (i.e., mean and variance) of the volume response variables (more details on study data are given in Section 2). The proposed model-based sampling method is motivated by the use of Gaussian processes [31] in sensor network optimization (SNO) problems [32–34]. The SNO problems are very similar to those in forest inventory sampling, since they both have the same common challenge, i.e., to optimize sample locations with respect to accuracy and cost efficiency. This is the main basis on which our proposed method is built upon, and for which Bayesian optimization with Gaussian processes offers a theoretical framework. Instead of using response and auxiliary data per se for new sampling decisions, our model-based approach lets the prediction model itself guide the sample selection using auxiliary (e.g., RS) data to areas where the model's prediction uncertainty for response (i.e., the inventory data) variables is the largest. By doing so, the proposed method aims to avoid similar data points to be included into the sample data, much alike as in the local pivotal method [24]. Since our approach relies on the prediction model, it is assumed that prior data exists for fitting the model before the sampling can be implemented. These data should be reflect the functional link between the auxiliary and response data as accurately as possible.

In general, NFIs typically cover hundreds of variables, of which information is recorded on NFI sample plots. These variables typically cover, among others, the trees (living or dead), site type, forest health as well as variables related to biodiversity or ecological

value [11,35]. However, it is not feasible to optimize the sampling design for all variables of interest. For example in Finnish NFI, the current systematic cluster sampling design is optimized for producing design unbiased estimates of the total volume of growing stock as well as the volumes of main tree species at regional and national level, and the same sampling design will be used also for all other variables recorded in NFI. In this study, we focus on the volumes of total growing stock and the main tree species groups due to their importance in national forest inventory.

Lastly, to summarize this study, our objective is to conduct empirical analysis on the performance of our proposed Bayesian model-based sampling method against two baseline methods in NFI data context. Our study data from Aland region consist of a population of NFI sample plots containing the response variables (volume of growing stock) and the auxiliary explanatory data consisting from airborne imagery and ALS data. We evaluate the performance of the sampling methods by measuring how accurate estimates they produce of the ground truth mean and variance population parameters of the volume response variables.

2. Materials

2.1. Study Area and Field Data

The real-world research data were collected from the archipelago province of Aland (lat. 60°11′ N, long. 20°22′ E) in Finland. The data set consisted from a set of airborne laser scanning (ALS), aerial imagery and reference data measured in the field. The ALS and aerial imagery data were used as predictor/input data to generalize the field reference data, i.e., response data, over a larger area. The total area covered by both ALS and aerial imagery data were approximately 346,000 ha, but a large part of it was sea area. The field data were mainly composed of 11th Finnish national forest inventory (NFI11) sample plots allocated on the basis of systematic cluster sampling. In the study area sample plot clusters were established in a grid of 3 × 3 km, and each cluster consisted of 9 sample plots in L-shaped form, having 200 m distance between plots (see Figure 1). In addition to these sample plot clusters, permanent clusters established in 9th NFI [35] were remeasured. A total of 349 NFI plots were measured in forestry land based on systematic sampling. Forestry land consists of forest land, other wooded land and unproductive land (see Figure 1 for map of land classes). In this study, we refer to a subset of the population that accurately reflects the characteristics of the larger group as a representative sample.

For RS-based forest inventory it is necessary to have field observations of all types of forest, otherwise the forest strata that are missing from field observations will be missing also from inventory results. In order to check the representativeness of the systematic sample, the inventory area was stratified into 196 strata based on ALS and aerial image features, and additional field observations were allocated to those strata that were underrepresented or missing in the systematic sample. The basis of the stratification was to check the representativeness of the systematic sample in relation to following features characterizing the forest types: species dominance (coniferous vs. broadleaved), average stand density, dominant height (Lorey's height) of trees and the spatial order of the trees. For the selection of the additional sample, an initial grid of plots was generated to inventory area with a spacing of 100 × 100 m between the initial plots (these points included the center point locations of the systematic sample plots). The plots of the initial grid were stratified into 196 strata based on ALS and aerial image features. The following features were used in the stratification: height where 85% of LiDAR returns have accumulated (4 strata), ratio of canopy returns to all returns (3 strata), inverse distance moment of rasterized canopy height model (4 strata) and spectral average of aerial image near-infrared (NIR) band (4 strata). These features correlate, respectively, with stand dominant height, stand density, size and spatial organization of tree crowns, and the proportion of broadleaved trees of the total canopy coverage (a more detailed description of ALS and aerial image features is presented in the following section). The strata, whose area was less than 100 ha in the entire inventory area, were excluded from additional sample, since their area was not significant. Thus,

although the theoretical number of strata was 192, some feature combinations are unlikely to occur, and the number of those strata representing an area more than 100 ha was 101 in the entire inventory area. The representativeness of the systematic sample was examined in relation to these strata, and altogether 126 additional plots were allocated to the strata underrepresented or missing among the systematic sample, bringing the total number of field plots to 475. The additional sample plots were selected as a random sample from each underrepresented or missing stratum, and they were not clustered. Otherwise, they were measured in the same way as systematic sample plots.

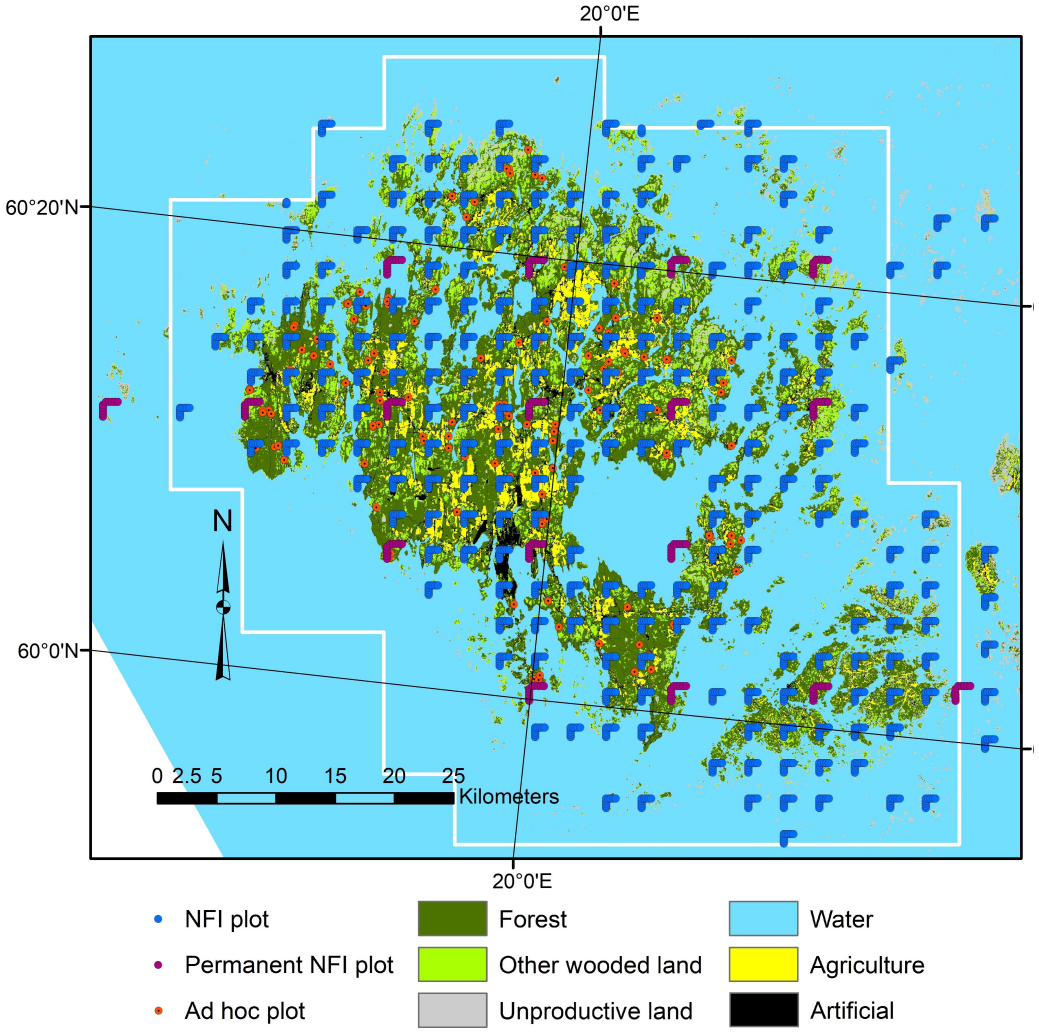

Figure 1. Map of the sampling layout in Aland. Background land cover map is based on NFI and topographic data provided by LUKE. The ALS coverage area is marked with white borderline.

The sample plots were measured as restricted relascope plots with a basal area (m^2/ha) factor 1 and maximum radius 9 m. For each sample plot, tree and stand level variables were recorded in accordance with NFI field guide and nomenclature [36]. The field variables that were applied for testing different sampling strategies in this study were volume of total growing stock and volumes per following tree species groups: pine, spruce and broadleaf trees. In practical forest inventories the amount of growing stock and proportions of tree species are typically the most important stand variables, especially for forest management [37].

2.2. Remote Sensing Data

The ALS and aerial imagery data contained a total of 154 variables covering point cloud features from ALS data as well as spectral and textural features from aerial imagery. The following features were extracted from ALS point cloud data from an area representing each 9 m radius sample plot [38–40]:

1. Average, standard deviation and coefficient of variation of height above ground (H) for canopy returns, separately from first (f) and last (l) returns (havg[f/l], hstd[f/l], hcv[f/l]).
2. H at which p% of cumulative sum of H of canopy returns is achieved (H_p) (hp[f/l], p is one of 0, 5, 10, 20, 30, 40, 50, 60, 70, 80, 85, 90, 95 and 100).
3. Percentage of canopy returns having $H \geq$ than corresponding H_p (p[f/l], p is one of 20, 40, 60, 80, 95).
4. Canopy densities corresponding to the proportions of points above fraction no. 0, 1, . . . , 9 to a total number of points (d0, d1, . . . , d9).
5. (a) Ratio of first canopy returns to all first returns (vegf), and (b) Ratio of last canopy returns to all last returns (vegl).
6. Ratio of intensity percentile p to the median of intensity for canopy returns (ip[f/l], p is one of 20, 40, 60 and 80).

The following features were extracted from the aerial image bands from an area representing the size of sample plots:

1. Average, standard deviation (std) and coefficient of variation (cv) from each of the four image bands: near-infrared (nir), red (r), green (g), blue (b).
2. The following multiband transformations Normalized difference vegetation index, NDVI. See, e.g., [41]: NDVI as (nir − r)/(nir + r), modified NDVI as (nir − g)/(nir + g), nir/r, nir/g.
3. Haralick textural features [42] based on co-occurrence matrices of image band values: angular second moment (ASM), contrast (Contr), correlation (Corr), variance (Var), inverse difference moment (IDM), sum average (SA), sum variance (SV), sum entropy (SE), entropy (Entr), difference variance (DV), difference entropy (DE).

Additionally, height and intensity values of LiDAR points were interpolated into raster format data with similar resolution as aerial imagery for extracting the same textural features as from aerial imagery. A detailed description of the acquisition of RS data is presented in [43,44].

The use of RS data in forest inventory estimation and survey sampling is widely studied, and related works can be found, e.g., in [44–48].

2.3. Target Population

In this research, we use the ten best features discovered from the ALS and aerial imagery data in the study [44] as the auxiliary data for the target response variables: volume of growing stock (all trees, pine, spruce and broadleaf) in the field reference data. In the referenced study, approximately ten predictor features were found to be sufficient to achieve optimal prediction performance for the corresponding response variables. We have listed the target response variables and their corresponding top ten used predictor features in Table 1. Histograms describing the value distributions of the target response variables is presented in Figure 2. The total number of available data points (i.e., our empirical study population) used in this study was 475. The goal is to estimate the mean and variance population parameters of the four target volume variables.

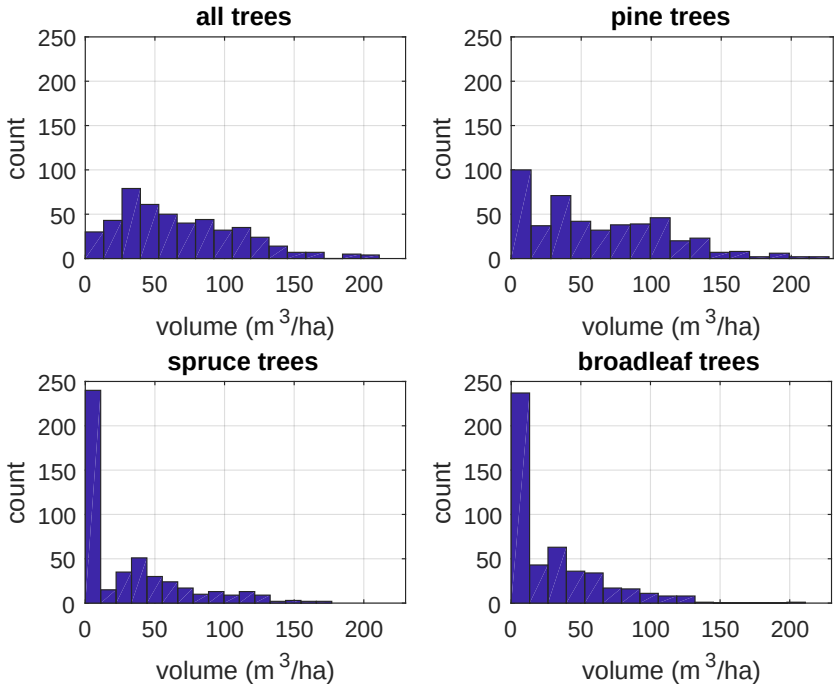

Figure 2. Histograms describing tree volume distributions in the Aland research area. Volumes of 0–10 m^3/ha of spruce and broadleaf trees are of high frequency in sample sites, whereas the most frequent volume of all trees in sample sites is 20–30 m^3/ha. The average volumes for all, pine, spruce and broadleaf trees are 67.4, 61.5, 30.5 and 26.7 m^3/ha respectively.

Table 1. List of the top ten predictor features found for the target variables: growing stock tree volume (all, pine, spruce, broadleaf trees) in the work by [44]. In the table H stands for height above ground. The optimal features were selected via greedy forward selection method see, e.g., [49].

Volume All Trees
texture feature, sum average, ALS based canopy height
percentage of last canopy returns above 20% height limit
percentage of first canopy returns above 90% height limit
texture feature, entropy, ALS based canopy height
texture feature, angular second moment, ALS based canopy height
texture feature, inverse difference moment, ALS based canopy height
H at which 100% of cumulative sum of last canopy returns is achieved (H_p, p%)
gndvi, transformation from band averages within the pixel windows: nir − g/nir + g
percentage of last canopy returns having $H \geq$ than corresponding H_{20}
coefficient of determination of first returned canopy returns
Volume Pine Trees
percentage of last canopy returns above 70% height limit
texture feature, angular second moment, ALS based intensity
texture feature, contrast, near-infrared band of CIR imagery
transformation from band averages within the pixel windows: nir/r
texture feature, sum average, ALS based canopy height
texture feature, sum average, ALS based intensity
texture feature, difference variance, blue band of RGB imagery

Table 1. *Cont.*

texture feature, difference entropy, near-infrared band of CIR imagery
ratio of last canopy returns to all last returns
ratio of first canopy returns to all first returns
Volume Spruce Trees
ratio of intensity percentile 20 to the median of intensity for last canopy returns
percentage of last canopy returns above 30% height limit
ratio of intensity percentile 60 to the median of intensity for last canopy returns
ratio of intensity percentile 80 to the median of intensity for first canopy returns
texture feature, difference variance, blue band of RGB imagery
texture feature, coefficient of determination, near-infrared band of CIR imagery
percentage of first canopy returns above 30% height limit
coefficient of determination of last returned canopy returns
texture feature, coefficient of determination, red band of CIR imagery
texture feature, contrast, blue band of RGB imagery
Volume Broadleaf Trees
texture feature, sum average, ALS based intensity
ratio of intensity percentile 40 to the median of intensity for first canopy returns
ratio of intensity percentile 20 to the median of intensity for first canopy returns
ratio of intensity percentile 40 to the median of intensity for last canopy returns
texture feature, entropy, ALS based intensity
texture feature, variance, ALS based intensity
texture feature, inverse difference moment, ALS based intensity
percentage of first canopy returns having $H \geq$ than corresponding H_{95}
percentage of first canopy returns above 20% height limit
H at which 5% of cumulative sum of last canopy returns is achieved (H_p, p%)

3. Methods

The following notation will be used throughout this and the following sections. A single observation of input or auxiliary predictor variables is denoted as a vector $\mathbf{x} \in \mathbb{R}^m$ with m distinct features. A corresponding response variable is denoted as $y \in \mathbb{R}$. The pair $\mathbf{d} = (\mathbf{x}, y)$ is treated as a single data point. For example, $\mathbf{x}$ might contain RS data (e.g., raster pixel information or derived features) on some geographic location and y could contain the average volume of trees in that corresponding location. In this study, $\mathbf{x}$ and y refer to the auxiliary and response data (forest inventory) as presented in earlier sections. An observed data set is denoted as $\mathcal{D} = (\mathcal{X}, \mathcal{Y})$, where $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ is the set of n input vectors and $\mathcal{Y} = \{y_1, y_2, \ldots, y_n\}$ is the set of n realizations of the response variable. A prediction model is denoted as $f(\mathbf{x}; \boldsymbol{\theta})$ (f in short), where $\boldsymbol{\theta} \in \mathbb{R}^q$ is a vector of model parameters. We use the symbol $\mathcal{U}$ to denote the population of all possible data points (i.e., all the data that can be sampled) and π_i ($0 \leq \pi_i \leq 1$) to denote the inclusion probability of data point $\mathbf{d}_i$. That is, π_i is the probability that the i^{th} data point of population $\mathcal{U}$ will be included into the observed (sampled) data set $\mathcal{D} \subset \mathcal{U}$. In other words, it is its probability of becoming part of the sample during the drawing of a single sample.

3.1. Population Parameter Estimators

In this study, we use the following simple sample mean and variance $(\hat{\mu}, \hat{\sigma^2})$ for estimating the ground-truth mean and variance (μ, σ^2) study population parameters of the four volume response variables:

$$\hat{\mu} = |\mathcal{U}|^{-1} \left(\sum_{\mathbf{d} \in \mathcal{D}} y + \sum_{\mathbf{d} \in \mathcal{U} \setminus \mathcal{D}} f(\mathbf{x}; \boldsymbol{\theta}) \right) \tag{1}$$

$$\hat{\sigma^2} = (|\mathcal{U}| - 1)^{-1} \left(\sum_{\mathbf{d} \in \mathcal{D}} (y - \hat{\mu})^2 + \sum_{\mathbf{d} \in \mathcal{U} \setminus \mathcal{D}} (f(\mathbf{x}; \boldsymbol{\theta}) - \hat{\mu})^2 \right), \tag{2}$$

where the unsampled response data ($\mathbf{d} \in \mathcal{U} \setminus \mathcal{D}$) are estimated by the model $f(\mathbf{x}; \boldsymbol{\theta})$. We compare the quality of sampling methods via these estimators empirically using mean squared errors (MSE) $E\left[(\hat{\mu} - \mu)^2\right]$ and $E\left[(\hat{\sigma^2} - \sigma^2)^2\right]$ where we take the averages with respect to the number of empirical experiments we conduct. We go into more details on this in Section 3.5.

The motivation of the proposed sampling method is that it attempts to include into the sample $\mathcal{D}$ such observations of the response variable y for which the prediction model $f(\mathbf{x}; \boldsymbol{\theta})$ has very high estimation uncertainty, given the auxiliary data $\mathbf{x}$. In other words, the proposed method aims to select such a sample $\mathcal{D}$ from the population $\mathcal{U}$, so that the estimation of the unsampled response data (the rightmost sum terms inside the parentheses in Equations (1) and (2)) would be as accurate as possible. That is, we do not want to include data into the sample $\mathcal{D}$ which yields no or small further gain (small model prediction uncertainty) for estimating the unsampled data ($\mathbf{d} \in \mathcal{U} \setminus \mathcal{D}$) using model $f(\mathbf{x}; \boldsymbol{\theta})$, but instead, data which yields most further gain (high model prediction uncertainty) in the estimation of the unsampled data using model $f(\mathbf{x}; \boldsymbol{\theta})$.

3.2. Simple Random and Local Pivotal Method Sampling

We will compare our proposed sampling algorithm with two baseline sampling methods: simple random sampling (SRS) [2] and local pivotal method sampling (LPM) [19,20,50]. In the SRS method, the inclusion probabilities for all data points are equal, i.e., $\pi_i = \pi_j \; \forall \, i, j$. In other words, the SRS samples data randomly with all the data points having equal probability of being included into the sample. In addition, the data points are sampled independently from one another.

The LPM is a sampling method based on the idea of avoiding the selection of data points that are similar in the feature space $\mathcal{X}$. The point is to select a spatially balanced data $\mathcal{D}$ from the population $\mathcal{U}$. LPM attempts to select the spatially balanced samples by locally aggregating the inclusion probabilities of neighboring data points, decreasing the likelihood that similar data samples are selected. This for example, is especially useful when we want to acquire a representative sample of geographical data. The LPM starts with an initial inclusion probability set $\Pi = \{\pi_1, \pi_2, \ldots, \pi_{|\mathcal{U}|}\}$ and proceeds by iteratively updating pairs of inclusion probabilities (π_i, π_j), so that the sampling outcome is decided for at least one of the two corresponding data points in each iteration.

This means that all the sampling decisions will be completed in at most $|\mathcal{U}|$ iterations of the algorithm. Note that in LPM, it is not required that $\pi_i = \pi_j \; \forall \, i, j$ but it is required that $\sum_{\mathbf{d}_i \in \mathcal{U}} \pi_i = n$, where n is the size of sampled data set $\mathcal{D}$ [20]. The main steps of the LPM sampling are the following:

1. Randomly choose a data point $\mathbf{d}_i \in \mathcal{U}$ with uniform probability.
2. Find the nearest neighbor (i.e., nearest in, e.g., Euclidean distance e sense) $\mathbf{d}_i$ of $\mathbf{d}_j$ in the feature space $\mathcal{X}$.
3. If data point $\mathbf{d}_i$ has two neighbors equally close in the feature space, then randomly with equal probability select either of the two neighbors.

4. Update the inclusion probability pair (π_i, π_j) using the rules found in Algorithm 1.
5. Remove the data point in the pair $(\mathbf{d}_i, \mathbf{d}_j)$ for which the inclusion probability is either 0 or 1 from further consideration.
6. If all the inclusion probabilities in set Π have $\pi_k = 1$ or $\pi_k = 0$, then stop the algorithm and include data points with $\pi_k = 1$ into $\mathcal{D}$. Otherwise, repeat from step 1.

The corresponding pseudocode for LPM is shown in Algorithm 1.

Algorithm 1 Pseudocode for LPM

Require: $\mathcal{U}, \Pi$ ▷ The population data and set of initial inclusion probabilities
Ensure: $\mathcal{D}$ ▷ The returned sample data
set $\mathcal{D} = \emptyset$ and $\mathcal{U}^* = \mathcal{U}$
while $|\mathcal{U}| > 0$ **do** ▷ Repeat until sampling decision is made for all the data
  Randomly select a data point $\mathbf{d}_i$ from set $\mathcal{U}$ with uniform probability
  Set $\mathbf{d}_j = \text{argmin}_{\mathbf{d} \in \mathcal{U} \setminus \mathbf{d}_i} e(\mathbf{x}, \mathbf{x}_i)$ ▷ find the nearest neighbor
  Update the inclusion probabilities $\pi_i, \pi_j \in \Pi$ using the rules:

$$\text{If } \pi_i + \pi_j < 1, \text{ then } (\pi_i, \pi_j) = \begin{cases} (0, \pi_i + \pi_j) & \text{with probability } \frac{\pi_j}{\pi_i + \pi_j} \\ (\pi_i + \pi_j, 0) & \text{with probability } \frac{\pi_i}{\pi_i + \pi_j}, \end{cases}$$

$$\text{else if } \pi_i + \pi_j \geq 1, \text{ then } (\pi_i, \pi_j) = \begin{cases} (1, \pi_i + \pi_j - 1) & \text{with probability } \frac{1-\pi_j}{2-\pi_i-\pi_j} \\ (\pi_i + \pi_j - 1, 1) & \text{with probability } \frac{1-\pi_i}{2-\pi_i-\pi_j}. \end{cases}$$

  Set $\mathcal{U} = \mathcal{U} \setminus \{\mathbf{d}_k \in \mathcal{U} : \pi_k \in \{0, 1\}\}$ ▷ Remove samples with decision
end while
Set $\mathcal{D} = \{\mathbf{d}_k \in \mathcal{U}^* : \pi_k = 1\}$ ▷ Data points with positive sampling decision
return $\mathcal{D}$

3.3. Data Sampling Via Bayesian Optimization

The sampling method we propose is based on utilizing a prediction model's uncertainty on the value of response variable y under a given input datum $\mathbf{x}$. To give motivation for the proposed method, we note that typically we have observations of the response variable y (such as forest growing stock) only in sampled points, whereas auxiliary data variables $\mathbf{x}$ (e.g., satellite/airborne data) are often available throughout the entire inventory area. This may often be the case in inventories that use RS-based auxiliaries such as the Finnish multi-source NFI [11]. We aim to utilize the link between the response variable and auxiliary information by firstly building a probabilistic model using the observed data set $\mathcal{D}$, and then basing the sampling decision on the model's conditional uncertainty on the value of y (quantified by its variance) given input feature datum $\mathbf{x}$. A new sample point is to be chosen based on where the prediction model has the highest uncertainty on the value of y. Whereas the sampling decisions with methods such as SRS or LPM focus mainly to variables y and $\mathbf{x}$ in itself, the sampling decisions in the proposed method are based on the functional link $y = f(\mathbf{x}; \boldsymbol{\theta}) + \epsilon$, where ϵ is Gaussian noise. The proposed method thus assumes that there exists strong enough correlation between the predictor variables $\mathbf{x}$ and the response variable y in order to utilize this link in data sampling. We will next go through the proposed method in a more detailed manner. Most of the following is based on literature by, e.g., [51–55]. Furthermore, more related literature based on Bayesian optimization can be found, e.g., in the works of [31,56–63].

Having observed a data set $\mathcal{D} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)\}$ of identically and independently distributed samples, we are interested in knowing the conditional distribution of y given a new input vector $\mathbf{x}_{\text{new}}$ and the data set $\mathcal{D}$. Explicitly put, we want to find out $p(y|\mathbf{x}_{\text{new}}, \mathcal{D})$, which can be written as:

$$p(y|\mathbf{x}_{\text{new}}, \mathcal{D}) = \int_{\mathbb{R}^q} p(y, \boldsymbol{\theta}|\mathbf{x}_{\text{new}}, \mathcal{D})\, d\boldsymbol{\theta} = \int_{\mathbb{R}^q} p(y|\mathbf{x}_{\text{new}}, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathcal{D})\, d\boldsymbol{\theta}, \quad (3)$$

where $p(\boldsymbol{\theta}|\mathcal{D})$ is the posterior distribution of model parameters. Note also that $p(y|\mathbf{x}_{\text{new}}, \boldsymbol{\theta}) = p(y|\mathbf{x}_{\text{new}}, \boldsymbol{\theta}, \mathcal{D})$. This follows from the fact that the model parameters $\boldsymbol{\theta}$ and $\mathbf{x}_{\text{new}}$ completely determine the distribution of y once the data $\mathcal{D}$ has been observed. Using Equation (3), we can now state the main statistic of interest in the proposed sampling method, which is the variance of the distribution $p(y|\mathbf{x}, \mathcal{D})$, i.e.,:

$$\sigma^2_{\mathcal{D}}(\mathbf{x}) = E_y\Big[(y-\mu)^2|\mathbf{x}, \mathcal{D}\Big], \tag{4}$$

where μ is the mean value of y w.r.t. distribution $p(y|\mathbf{x}, \mathcal{D})$ and E_y stands for expectation w.r.t. same distribution. We see from Equation (4) that the variance is a function of $\mathbf{x}$, but not $\mathcal{D}$ since we assume this to be fixed. In this study, we call the proposed sampling method (based on the statistic in Equation (4)) *Bayesian maximum variance inclusion* (BMVI). The BMVI always chooses sample data points $\mathbf{d} = (\mathbf{x}, y)$ where $\sigma^2_{\mathcal{D}}(\mathbf{x})$ attains highest values (i.e., maximum uncertainty). The pseudocode for the BMVI is illustrated in Algorithm 2 and corresponding process flow chart is presented in Figure 3. The symbols $k, \mathcal{D}_p, \mathcal{D}_s$ denote the number of data points to be sampled, a prior data set available for calculating the posterior predictive distribution $p(y|\mathbf{x}, \mathcal{D}_p)$, and the new sampled data set (i.e., $k = |\mathcal{D}_s|$). The algorithm shows that the inclusion probabilities are $\pi_i = 1$ for the k single data samples with the highest posterior predictive variances. For all the remaining data points the inclusion probabilities are $\pi_i = 0$.

Algorithm 2 Pseudocode for BMVI

Require: $\mathcal{D}_p, \mathcal{U}, k$ ▷ Prior data set, population and sample size
Ensure: $\mathcal{D}_s$ ▷ Sample data set
1: Set $\mathcal{D}_s = \emptyset$
2: Calculate $p(y|\mathbf{x}, \mathcal{D}_p)$ using prior data set $\mathcal{D}_p$ ▷ Note $\mathcal{D}_p \subset \mathcal{U}$
3: **for** $i \leftarrow 1$ to k **do** ▷ Select k data points
4: Set $(\mathbf{x}_i, y_i) = \text{argmax}_{(\mathbf{x},y)\in\mathcal{U}\setminus\mathcal{D}_p} \sigma^2_{\mathcal{D}}(\mathbf{x})$ ▷ Data point with max. uncertainty
5: Set $\mathcal{D}_s = \mathcal{D}_s \cup \{(\mathbf{x}_i, y_i)\}$ ▷ Include data point into sample
6: Set $\mathcal{U} = \mathcal{U} \setminus \{(\mathbf{x}_i, y_i)\}$ ▷ Remove sampled point from population
7: **end for**
8: **return** $\mathcal{D}_s$ ▷ Return sample of size k

After making Gaussian assumptions on the distributions in Equation (3), it follows that the variance statistic of Equation (4) can be written as:

$$\sigma^2_{\mathcal{D}}(\mathbf{x}) = \frac{1}{\beta} + \mathbf{g}(\mathbf{x})^T \mathbf{A}^{-1} \mathbf{g}(\mathbf{x}), \tag{5}$$

where $\beta > 0$ is a parameter controlling the prior variance of the response variable y, $\mathbf{g}$ is a gradient vector of the prediction model $f(\mathbf{x}; \boldsymbol{\theta})$ evaluated at a maximum posterior point, and $\mathbf{A}$ is the Hessian matrix of the exponent of the posterior distribution of model weights $\boldsymbol{\theta}$. Detailed definitions and derivations of this result can be found from the Appendix A part of this study. A Python implementation of the proposed BMVI method and example demonstration made by the authors of this study can be found from [64].

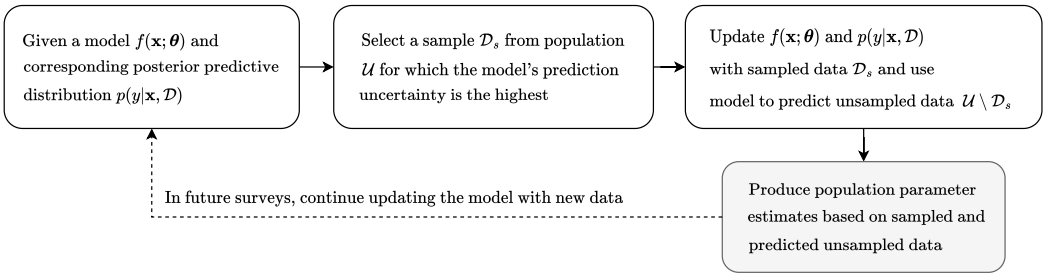

Figure 3. Process flow chart of the proposed BMVI sampling method.

3.4. Prediction Models

Next, we will give a short introduction to the prediction models $f(\mathbf{x};\boldsymbol{\theta})$ we apply in the empirical analyses of the following sections.

3.4.1. Ridge Regression

The first prediction method used in our analyses is ridge regression known also as regularized least squares (RLS) [65]. RLS is almost identical to basic linear regression method, with the exception that instead of minimizing simply the squared error between observed data and predictions, the RLS adds a regularizing term into the squared error minimization. This addition makes the model selection process to favor more well-behaving models, which are more likely to achieve successful generalization to new unseen data, e.g., [66]. Explicitly, in RLS the prediction model is simply a linear function of the input data, i.e., $f(\mathbf{x};\boldsymbol{\theta}) = \boldsymbol{\theta}^T\mathbf{x} + \theta_0$ where θ_0 denotes the constant bias term of the model. In RLS, the model parameters are selected so as to minimize the (error) function:

$$S(\boldsymbol{\theta}) = \frac{\beta}{2}\sum_{i=1}^{n}\{y_i - \boldsymbol{\theta}^T\mathbf{x}_i - \theta_0\}^2 + \frac{\alpha}{2}\sum_{j=1}^{m}\theta_j^2, \tag{6}$$

where $\alpha, \beta > 0$ and α controls the degree of regularization. The constants α, β correspond directly to those in Equations (A1), (A2) and (A4) (see Appendix A), showing the connection between Tikhonov regularization and Bayesian modelling [67]. Note that it is not necessary to include the constant term θ_0 into the second term in Equation (6) since it simply controls the offset of the hyperplane $f(\mathbf{x};\boldsymbol{\theta})$ but not its slopes. In our analyses, the RLS hyperparameter selection (i.e., α, β) was conducted using leave-one-out cross-validation (LOOCV) [68].

3.4.2. Multilayer Perceptron

In addition to the RLS, a multilayer perceptron (MLP) [51] was tested as a prediction model. A MLP is a feedforward neural network defined by the number of hidden layers L, inputs and outputs, hidden nodes and types of activation functions, and it has shown great performance in a number of applications. The MLP network is trained by minimizing a suitable error function, such as $S(\boldsymbol{\theta})$ in the Equation (6). The parameters of a MLP can be defined as the set:

$$\boldsymbol{\theta} \equiv \left\{\theta_{ij}^{(l)} \mid 1 \le l \le L+1, 0 \le i \le d^{(l-1)}, 1 \le j \le d^{(l)}\right\}, \tag{7}$$

where $d^{(l)}$ is the number of nodes on layer l. In other words, $\theta_{ij}^{(l)}$ means a network weight connecting node i at layer $l-1$ to node j at layer l. The weights $\theta_{ij}^{(1)}$ and $\theta_{ij}^{(L+1)}$ correspond to weights connected to the input and output nodes respectively. As an example, a MLP with one hidden layer ($L = 1$) can be explicitly expressed as a function:

$$f(\mathbf{x};\boldsymbol{\theta}) = f_2\left(\sum_{j=1}^{d^{(1)}}\theta_{j1}^{(2)} f_1\left(\sum_{i=1}^{m}\theta_{ij}^{(1)}x_i\right)\right), \tag{8}$$

where now weights $\theta_{ij}^{(1)}$ and $\theta_{j1}^{(2)}$ correspond to connections of the hidden layer to input and output layers correspondingly. The functions $f_1(\cdot)$ and $f_2(\cdot)$ correspond to the activation functions, which need not be the same at all layers. Common choices for the activation functions are, e.g., linear or sigmoid functions. In our experiments, we used a MLP model provided by the NETLAB-library [69]. The MLP network was trained using the scaled conjugate gradient algorithm [70].

3.5. Implementation Details of the Empirical Analysis

Lastly, in this section we will describe the technical details of the empirical analyses in order to make it more clear on how to interpret the results of the next section. The corresponding results in Tables 2 and 3 were produced using the Algorithm 3 presented in this section. Note that the emphasis of this study was not to find an optimal prediction model (e.g., the RLS or MLP) for the data sets, but the comparison of the sampling methods by their performance in the estimation of response variable population parameters. Thus due to their irrelevance, no optimal prediction model parameters are listed in this study. Recall, that we denoted the data population as $\mathcal{U}$ and the prediction model as f. In addition, we will denote a sampling method as S_M, i.e., $S_M \in \{\text{SRS, LPM, BMVI}\}$, and sample data sets as $\mathcal{D}_p, \mathcal{D}_s \subset \mathcal{U}$ where we have $\mathcal{D}_p \cap \mathcal{D}_s = \emptyset$.

Table 2. Results of population mean μ estimations in terms of MSE. The results are illustrated for all response variables, sampling methods and prediction models. The leftmost column of the table represents different valued sampling fraction vector **f** introduced in the methods Section 3.5. In each group of three (SRS, LPM, BMVI) the best sampling method is emphasized with a bolded MSE value. The response variables for volume of growing stock (all trees, pine trees, spruce trees, broadleaf trees) are denoted as v_a, v_p, v_s and v_b respectively.

		Regularized Least Squares				**Multilayer Perceptron**			
		v_a	v_p	v_s	v_b	v_a	v_p	v_s	v_b
0.1/0.6/0.3	SRS	1.510	2.829	1.794	**1.399**	**1.778**	3.176	**2.732**	**1.911**
	LPM	1.198	2.595	**1.382**	1.639	1.973	**2.361**	2.735	3.757
	BMVI	**0.832**	**2.036**	20.065	6.502	11.086	16.368	6.977	2.036
0.2/0.5/0.3	SRS	1.353	3.072	1.642	1.583	1.913	2.349	2.180	2.333
	LPM	1.229	2.044	**1.307**	**1.384**	**1.585**	**2.181**	2.276	2.202
	BMVI	**0.796**	**1.430**	19.881	6.796	7.030	16.576	**2.054**	**1.561**
0.3/0.4/0.3	SRS	1.536	2.502	2.111	**1.731**	**2.323**	**2.416**	2.047	2.226
	LPM	1.733	2.589	**1.851**	1.747	2.610	2.439	2.303	2.774
	BMVI	**1.046**	**2.158**	17.148	5.848	11.222	9.252	**1.800**	**1.659**
0.4/0.3/0.3	SRS	1.381	2.083	2.085	1.579	**1.935**	2.077	2.612	2.378
	LPM	1.241	2.416	**1.648**	**1.414**	2.006	**1.852**	2.472	2.248
	BMVI	**0.945**	**2.037**	12.111	6.148	3.024	5.594	**2.332**	**1.853**
0.5/0.2/0.3	SRS	1.588	2.434	1.812	**1.623**	1.851	2.638	2.176	2.804
	LPM	1.184	**2.195**	**1.569**	1.853	**1.472**	**1.955**	2.638	2.104
	BMVI	**0.871**	2.321	6.851	3.495	2.621	3.756	**1.443**	**1.015**
0.6/0.1/0.3	SRS	1.211	3.615	**1.613**	1.861	**1.824**	2.811	1.795	2.491
	LPM	1.184	2.996	1.830	**1.353**	2.494	**2.341**	2.186	2.059
	BMVI	**1.094**	**1.641**	2.761	1.855	2.782	4.221	**0.844**	**0.861**

Table 3. Analogous results as in Table 2 but for population variance σ^2 estimations. All the values in table below have a common factor of 10^5 which has been omitted from the values for clearer presentation. The response variables for volume of growing stock (all trees, pine trees, spruce trees, broadleaf trees) are denoted as v_a, v_p, v_s and v_b respectively.

		Regularized Least Squares				**Multilayer Perceptron**			
		v_a	v_p	v_s	v_b	v_a	v_p	v_s	v_b
0.1/0.6/0.3	SRS	6.255	8.666	8.204	6.632	**7.555**	**4.972**	3.799	3.726
	LPM	7.829	9.072	8.395	5.202	9.882	6.821	3.323	4.655
	BMVI	**0.243**	**5.499**	**1.244**	**0.699**	12.202	9.341	**2.962**	**1.954**
0.2/0.5/0.3	SRS	7.293	8.661	9.115	6.638	9.562	5.906	3.606	4.271
	LPM	6.981	9.381	8.197	5.818	9.950	**5.438**	3.460	3.082
	BMVI	**0.184**	**5.184**	**0.892**	**0.710**	**6.899**	7.962	**0.379**	**0.997**
0.3/0.4/0.3	SRS	5.697	8.377	9.399	5.952	8.268	5.231	3.753	3.758
	LPM	8.497	8.346	9.627	5.682	9.737	5.789	4.097	4.362
	BMVI	**0.258**	**5.513**	**0.935**	**0.704**	**4.242**	**3.499**	**0.180**	**0.168**
0.4/0.3/0.3	SRS	4.365	9.241	9.490	5.282	7.489	5.902	4.165	2.660
	LPM	5.726	7.925	8.725	5.088	6.747	5.516	3.147	2.730
	BMVI	**0.436**	**5.255**	**2.078**	**0.683**	**3.936**	**3.598**	**0.388**	**0.247**
0.5/0.2/0.3	SRS	5.527	9.382	7.713	5.088	7.093	5.786	2.846	3.394
	LPM	5.282	8.416	8.212	6.579	8.525	5.609	3.486	3.807
	BMVI	**0.336**	**5.418**	**4.095**	**1.288**	**2.364**	**2.495**	**0.246**	**0.163**
0.6/0.1/0.3	SRS	5.263	8.452	7.706	5.826	10.022	5.101	3.254	3.384
	LPM	4.405	9.165	8.143	5.556	11.979	6.345	3.713	2.682
	BMVI	**0.469**	**6.016**	**5.155**	**1.743**	**1.698**	**2.948**	**0.223**	**0.553**

Algorithm 3 Procedure used for obtaining the empirical results

Require: $\mathcal{U}, \mathbf{f}, f$ ▷ Population data, sample fraction vector and prediction model
Ensure: $MSE_{\mu}^{\text{SRS}}, MSE_{\sigma^2}^{\text{SRS}}, MSE_{\mu}^{\text{LPM}}, MSE_{\sigma^2}^{\text{LPM}}, MSE_{\mu}^{\text{BMVI}}, MSE_{\sigma^2}^{\text{BMVI}}$
Set $SE_{\mu}^{\text{SRS}} = \emptyset,\ SE_{\mu}^{\text{LPM}} = \emptyset,\ SE_{\mu}^{\text{BMVI}} = \emptyset$ ▷ Sets of squared error values
Set $SE_{\sigma^2}^{\text{SRS}} = \emptyset,\ SE_{\sigma^2}^{\text{LPM}} = \emptyset,\ SE_{\sigma^2}^{\text{BMVI}} = \emptyset$
for $i \leftarrow 1$ to 100 **do** ▷ Repeat 100 times to produce averaged results
  Select a random prior sample set $\mathcal{D}_p$ from $\mathcal{U}$ according to $\mathbf{f}$
  for $S_M \in \{\text{SRS, LPM, BMVI}\}$ **do** ▷ Do sampling with all methods
    Select a sample $\mathcal{D}_s$ from $\mathcal{U} \setminus \mathcal{D}_p$ using $S_M, \mathbf{f}, f$ and $\mathcal{D}_p$
    Set $\mathcal{D} = \mathcal{D}_p \cup \mathcal{D}_s$ ▷ Combine prior and sampled data
    Set $V = \mathcal{U} \setminus \mathcal{D}$ ▷ Use the remaining unsampled data for testing
    Train a prediction model f using data set $\mathcal{D}$
    Set estimator $\hat{\mu} = |\mathcal{U}|^{-1}[\sum_{\mathbf{d}\in\mathcal{D}} y + \sum_{\mathbf{d}\in V} f(\mathbf{x};\boldsymbol{\theta})]$
    Set estimator $\hat{\sigma^2} = (|\mathcal{U}|-1)^{-1}\left[\sum_{\mathbf{d}\in\mathcal{D}}(y-\hat{\mu})^2 + \sum_{\mathbf{d}\in V}(f(\mathbf{x};\boldsymbol{\theta})-\hat{\mu})^2\right]$
    Set $SE_{\mu}^{S_M}[i] = (\hat{\mu}-\mu)^2$ ▷ Error between estimate and true value
    Set $SE_{\sigma^2}^{S_M}[i] = (\hat{\sigma^2}-\sigma^2)^2$
  end for
end for
for $S_M \in \{\text{SRS, LPM, BMVI}\}$ **do** ▷ Calculate MSEs for all methods
  Set $MSE_{\mu}^{S_M} = \text{mean}\left(SE_{\mu}^{S_M}\right)$
  Set $MSE_{\sigma^2}^{S_M} = \text{mean}\left(SE_{\sigma^2}^{S_M}\right)$
end for ▷ Lastly return all MSE values for all methods
return $MSE_{\mu}^{\text{SRS}}, MSE_{\sigma^2}^{\text{SRS}}, MSE_{\mu}^{\text{LPM}}, MSE_{\sigma^2}^{\text{LPM}}, MSE_{\mu}^{\text{BMVI}}, MSE_{\sigma^2}^{\text{BMVI}}$

Since the core principle behind the BMVI sampling method is in utilizing the learned functional link between $\mathbf{x}$ and y, the method assumes that we have some prior data set $\mathcal{D}_p$ available for training the model f before we conduct the sampling of new data, i.e.,

$\mathcal{D}_s$ via BMVI. In clearer terms, the BMVI uses previously sampled data to optimize future sampling decisions. Thus in the empirical experiments, it is always assumed that we have some prior data set $\mathcal{D}_p$ available before the actual sampling of $\mathcal{D}_s$ is conducted with given S_M. Furthermore, since the sampling decisions of the BMVI method are obviously affected by the data used for training the prediction model f, it is of interest to study how the size of the prior training data $\mathcal{D}_p$ with respect to the whole data population $\mathcal{U}$ affects the sampling performance of the BMVI. For this reason, we parameterize our experiments with a vector:

$$\mathbf{f} = \left(\frac{|\mathcal{D}_p|}{|\mathcal{U}|}, \frac{|\mathcal{D}_s|}{|\mathcal{U}|}, \frac{|\mathcal{U} \setminus (\mathcal{D}_p \cup \mathcal{D}_s)|}{|\mathcal{U}|} \right) \in (0,1)^3. \quad (9)$$

In other words, the elements of the vector $\mathbf{f}$ are interpreted as: (1) the fraction of data points of the population $\mathcal{U}$ available in the prior set $\mathcal{D}_p$, (2) the fraction of new data to be sampled into set $\mathcal{D}_s$ with a given sampling method S_M, and (3) the remaining fraction of the population data (i.e., out-of-sample data) used for testing the estimation performance of population parameters. In our experiments we used a reasonable constant fraction of 30% of the data for testing the estimation performance. Thus we always had in the experiments that $\frac{|\mathcal{D}_p|}{|\mathcal{U}|} + \frac{|\mathcal{D}_s|}{|\mathcal{U}|} = 0.7$, with $\frac{|\mathcal{D}_p|}{|\mathcal{U}|}, \frac{|\mathcal{D}_s|}{|\mathcal{U}|} \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6\}$. For readers concerned with the high fraction of prior data (e.g., 0.6), we by no means imply that one should obtain 60% of data population prior to implementing the sampling. The high prior data fraction is simply added into the study to illustrate how the performance of the BMVI changes as more data are available for model training. This allows the BMVI method to, e.g., utilize sampled data from previous years in future sampling in the Bayesian paradigm manner.

The complete procedure used for obtaining the results of the next section is described in Algorithm 3. The algorithm is parametrized by the used data set $\mathcal{U}$, a fraction vector $\mathbf{f}$ and a prediction model f. The algorithm returns for all three sampling methods S_M the MSE values of the population mean μ and variance σ^2 parameter estimations. Note on line 3 of the algorithm that we repeat the experiments 100 times. This is due to decrease the effect of randomness in the estimation statistics by providing averaged results. For guaranteeing a valid comparison in the analysis results, all the sampling methods (i.e., SRS, LPM, BMVI) shared the same prior data set $\mathcal{D}_p$ when implementing a single comparative calculation run (line 4 in the algorithm). In addition, note in line 6 that only the BMVI method is dependent on f and $\mathcal{D}_p$. Rest of the algorithm is straightforward and on lines 10–11 the population parameters are estimated using the data set $\mathcal{D} = \mathcal{D}_p \cup \mathcal{D}_s$ and the auxiliary $\mathbf{x}$ data available in the set V. Recall that all the auxiliary data $\mathbf{x}$ (i.e., RS data) is assumed to be fully known throughout the research area and the response variable data y (NFI volume of growing stock) is only partly known and requires further sampling.

The auxiliary predictor features used in the real-world data case are listed in Table 1. The response variables for volume of growing stock (all trees, pine trees, spruce trees, broadleaf trees) are denoted in the results in Tables 2 and 3 as v_a, v_p, v_s and v_b respectively.

4. Results

In this section, we will go through the empirical results of comparing the BMVI method with SRS and LPM sampling using the previously described data sets and prediction models. Refer to Algorithm 3 for technical details on the results.

Volume of Growing Stock Data

The empirical results of this study are illustrated in Tables 2 and 3. The leftmost column in the tables represent different valued vectors $\mathbf{f}$ (i.e., fraction of data in $\mathcal{D}_p, \mathcal{D}_s$ and in the unsampled data set V). For example the values 0.1/0.6/0.3 in the first row means that 10% of the data in population $\mathcal{U}$ is assumed to be known beforehand (as we must have some prior data for the BMVI), 60% of the data will be sampled from the population, and the remaining 30% is estimated using the observed data $\mathcal{D} = \mathcal{D}_p \cup \mathcal{D}_s$ and prediction model f.

In population mean μ estimation, we can see the BMVI performing best for response variables v_a (volume of all trees) and v_p (volume of pine trees) with a linear prediction model RLS. For response variables v_s (volume of spruce trees) and v_b (volume of broadleaf trees), best results are achieved by the BMVI with MLP prediction model. In the case of population variance σ^2 estimation, we see the BMVI performing best of the three sampling methods in almost all cases. According to the results, the BMVI method outperforms SRS and LPM by a significant amount in population variance estimation, which makes intuitive sense based on the design of the method: sample from locations with large prediction uncertainty (i.e., prediction variance). In the case of population mean estimation for v_s and v_b we can notice the effect of the used prediction model.

To summarize the results, we firstly notice that the performance of the proposed method seems to increase as the amount of prior training data available for the prediction model increases. This makes sense because the BMVI method relies on the model. Secondly, the used model has a clear effect on the performance of the proposed method. If the model is suitable, the BMVI method outperforms the baselines and vice versa. The performance of the baselines is more stable in this sense. Lastly, the population variance estimation is more effective than population mean estimation with the BMVI, understandably so, due to the design of the BMVI method.

5. Discussion

The benefits of using auxiliary data in forest inventory sampling has been noted in the corresponding literature, e.g., [21,24]. In the work of [24] the authors showed significant improvements in sampling efficiency for forest inventories with the usage of auxiliary remote sensing data. In addition, in [21] experiments made with both synthetic and real data showed great utility of using airborne laser scanning data in forest inventory sampling design. Furthermore, the application of Bayesian approaches in optimizing geostatistical sampling designs can be found from a variety of literature, e.g., in the works by [58–60,63].

Although forest inventories such as NFIs usually collect information of hundreds of variables, the sampling design is typically optimized for variables of primary importance, such as volumes per tree species (assuming that the sampling design is suitable also for other variables), which has been noted, e.g., in studies by [21,24] where tree volume has been the main variable of interest. The corresponding sampling design optimized for tree volume will also be used also for all other variables recorded in the NFI, since it is theoretically and practically impossible to optimize the sampling design for all variables of interest simultaneously.

In this work, the real-world NFI data were available at the time as a systematic cluster sample from the research area. The promising empirical results encourage the continued research of applying Bayesian optimization methods in the context of RS-based forest inventory sampling. The current version of the BMVI method offers deterministic sampling, since the field samples with maximum prediction variance are always selected and thus the method does not produce a probability sample [1]. However, the method can be easily altered into producing a probability sample by having the sample inclusion probabilities being proportional to prediction uncertainties. That is, field samples with highest prediction uncertainties have the highest inclusion probabilities.

Considering the practical work-flow of the proposed method in operational inventory procedures, the method requires that we have some field samples already available since the prediction model needs to be trained before the BMVI sampling can be applied. This requirement is in the core of the method, since it uses previous data to decide where to place new field samples. Other practical challenges might occur if the method suggests field samples in geographical areas which are difficult to access, though all sampling methods are limited by these areas.

The advantages of the proposed method is that it allows the updating of the corresponding prediction model using new data with the expectation that the method's performance increases as new data are received, which the empirical results of this study

support. Other advantage of the method is that it allows, via the prediction model, to decrease redundant information being included (in the sense of similar sample points) into the sample data, which would yield of no or small further gain in the BMVI method's performance improvement. Disadvantages of the method include the susceptibility of the BMVI to model selection, which we have seen to have a clear effect on the method's performance. Furthermore, since the performance of the BMVI method also partly relies on utilizing new data for the model, it might be difficult to obtain useful new data for the BMVI due to the fact that the NFI often can use permanent sample plots for other reasons (e.g., tree growth measurements over time).

Regarding the results of this study, it is good to also note the volume distribution histograms described in Figure 2 and their effect. These histograms represent the distribution of the sample in the Aland region and thus the results do not automatically extrapolate into the whole of Finland where the corresponding tree distributions can be different.

6. Conclusions

In this study, we proposed a data sampling method based on Bayesian optimization, the BMVI, which utilizes the model link between the auxiliary RS and the NFI variables in new NFI field sample decision-making. We compared the BMVI against SRS and LPM sampling methods by measuring their performance in terms of MSE in producing estimates for NFI volume of growing stock population parameters, namely the mean and variance. The empirical results showed overall best performance for the proposed method when compared with the baselines, especially when enough training data (which we also called prior data earlier) was available to learn the model link between RS and forest inventory variables. The results also revealed the relevance of the underlying prediction model, which should be optimized based on the response variable of interest.

Author Contributions: Conceptualization, J.P. and J.H.; methodology, J.P. and J.H.; formal analysis, J.P.; investigation, J.P., S.T. and J.H.; resources, S.T.; writing—original draft J.P. and S.T.; writing—review and editing, J.P., S.T. and J.H.; supervision J.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The study data were provided by the Natural Resource Institute Finland (Luke) and are not publicly available on default. Further inquiries on the data should be directed to Luke. An example code of the proposed sampling method can be freely accessed in [64].

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Derivations and Proofs

In this appendix, we will derive the result in Equation (5) of the methods section. We will next proceed with formulating a closed-form expression for $\sigma^2_{\mathcal{D}}(\mathbf{x})$. To begin, we assume a Gaussian prior distribution for the prediction model parameters:

$$p(\boldsymbol{\theta}) \propto \exp\left(-\frac{\alpha}{2}||\boldsymbol{\theta}||^2\right) = \exp\left(-\frac{\alpha}{2}\sum_{j=1}^{q}\theta_j^2\right) = \prod_{j=1}^{q}\exp\left(-\frac{\alpha}{2}\theta_j^2\right). \tag{A1}$$

That is, each model parameter θ_j is assumed to be distributed as $\theta_j \sim \mathcal{N}(0, \alpha^{-1})$. The response variable y is assumed to be generated by a function $f(\mathbf{x}; \boldsymbol{\theta})$ with additive zero-mean Gaussian noise $\epsilon \sim \mathcal{N}(0, \beta^{-1})$, i.e.,:

$$p(y|\mathbf{x}, \boldsymbol{\theta}) \propto \exp\left(-\frac{\beta}{2}\epsilon^2\right), \tag{A2}$$

where $\epsilon = y - f(\mathbf{x};\boldsymbol{\theta})$. By also assuming that the data set $\mathcal{D}$ consists from n identically and independently distributed samples, we get the data likelihood as:

$$p(\mathcal{D}|\boldsymbol{\theta}) \propto \prod_{i=1}^{n} p(y|\mathbf{x}_i,\boldsymbol{\theta}) = \exp\left(-\frac{\beta}{2}\sum_{i=1}^{n}\{y_i - f(\mathbf{x}_i;\boldsymbol{\theta})\}^2\right). \tag{A3}$$

We can now use Equations (A1) and (A3) to express the posterior distribution for $\boldsymbol{\theta}$ as:

$$p(\boldsymbol{\theta}|\mathcal{D}) \propto p(\mathcal{D}|\boldsymbol{\theta})\, p(\boldsymbol{\theta}) \propto \exp\left(-\frac{\beta}{2}\sum_{i=1}^{n}\{y_i - f(\mathbf{x}_i;\boldsymbol{\theta})\}^2 - \frac{\alpha}{2}\sum_{j=1}^{q}\theta_j^2\right). \tag{A4}$$

Next, we will denote the negative of the exponent in Equation (A4) as

$$S(\boldsymbol{\theta}) = \frac{\beta}{2}\sum_{i=1}^{n}\{y_i - f(\mathbf{x}_i;\boldsymbol{\theta})\}^2 + \frac{\alpha}{2}\sum_{j=1}^{q}\theta_j^2, \tag{A5}$$

and make a second degree Taylor approximation for this function around the maximum posterior point $\boldsymbol{\theta}_{\mathrm{MP}} = \mathrm{argmax}_{\boldsymbol{\theta}\in\mathbb{R}^q}\, p(\boldsymbol{\theta}|\mathcal{D})$:

$$S(\boldsymbol{\theta}) \approx S(\boldsymbol{\theta}_{\mathrm{MP}}) + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_{\mathrm{MP}})^T\mathbf{A}(\boldsymbol{\theta} - \boldsymbol{\theta}_{\mathrm{MP}}), \tag{A6}$$

where $\mathbf{A}$ is the Hessian matrix of $S(\boldsymbol{\theta})$ evaluated at $\boldsymbol{\theta}_{\mathrm{MP}}$, i.e., the $(i,j)^{\mathrm{th}}$ element of $\mathbf{A}$ is

$$\mathbf{A}_{i,j} = \frac{\partial^2}{\partial\theta_i\partial\theta_j}(S(\boldsymbol{\theta}))|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\mathrm{MP}}}.$$

The maximum posterior $\boldsymbol{\theta}_{\mathrm{MP}}$ corresponds also to the parameters, which minimize $S(\boldsymbol{\theta})$, i.e., $\boldsymbol{\theta}_{\mathrm{MP}} = \mathrm{argmin}_{\boldsymbol{\theta}\in\mathbb{R}^q}\, S(\boldsymbol{\theta})$. Furthermore, note that the prior distribution $p(\boldsymbol{\theta})$ provides a regularizing function into $S(\boldsymbol{\theta})$ which results in favoring smaller values of θ_j, thus encouraging the selection of smoother functions $f(\mathbf{x};\boldsymbol{\theta})$ in $\boldsymbol{\theta}_{\mathrm{MP}}$ solution. Assuming in addition that the width of the posterior distribution of $\boldsymbol{\theta}$ is sufficiently narrow (due to the Hessian $\mathbf{A}$), we can approximate $f(\mathbf{x};\boldsymbol{\theta})$ with a linear expansion around $\boldsymbol{\theta}_{\mathrm{MP}}$ as $f(\mathbf{x};\boldsymbol{\theta}) \approx f(\mathbf{x};\boldsymbol{\theta}_{\mathrm{MP}}) + \mathbf{g}^T(\boldsymbol{\theta} - \boldsymbol{\theta}_{\mathrm{MP}})$, where $\mathbf{g}$ is the gradient vector of $f(\mathbf{x};\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ evaluated at $\boldsymbol{\theta}_{\mathrm{MP}}$. That is, the j^{th} element of $\mathbf{g}$ is:

$$g_j = \frac{\partial}{\partial\theta_j}(f(\mathbf{x};\boldsymbol{\theta}))|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\mathrm{MP}}}.$$

The linear approximation of $f(\mathbf{x};\boldsymbol{\theta})$ is suitable here without significantly losing accuracy, since most of the probability mass is focused on $\boldsymbol{\theta}_{\mathrm{MP}}$ and the higher order terms of the expansion are close to zero. By now plugging Equations (A2) and (A4) into Equation (3), using the approximations of $S(\boldsymbol{\theta})$ and $f(\mathbf{x};\boldsymbol{\theta})$ and denoting $\Delta\boldsymbol{\theta} = \boldsymbol{\theta} - \boldsymbol{\theta}_{\mathrm{MP}}$ and $\hat{y}_{\mathrm{MP}}(\mathbf{x}) = f(\mathbf{x};\boldsymbol{\theta}_{\mathrm{MP}})$, we get the expression for the posterior predictive distribution for y in Equation (3) as:

$$\begin{aligned} p(y|\mathbf{x},\mathcal{D}) &\propto \int_{\mathbb{R}^q} \exp\left(-\frac{\beta}{2}\{y - \hat{y}_{\mathrm{MP}}(\mathbf{x}) - \mathbf{g}^T\Delta\boldsymbol{\theta}\}^2\right) \\ &\quad\times \exp\left(-S(\boldsymbol{\theta}_{\mathrm{MP}}) - \frac{1}{2}\Delta\boldsymbol{\theta}^T\mathbf{A}\Delta\boldsymbol{\theta}\right) d\boldsymbol{\theta} \\ &\propto \int_{\mathbb{R}^q} \exp\left(-\frac{\beta}{2}\{y - \hat{y}_{\mathrm{MP}}(\mathbf{x}) - \mathbf{g}^T\Delta\boldsymbol{\theta}\}^2 - \frac{1}{2}\Delta\boldsymbol{\theta}^T\mathbf{A}\Delta\boldsymbol{\theta}\right) d\boldsymbol{\theta} \\ &= (2\pi)^{q/2}|\mathbf{A} + \beta\mathbf{g}\mathbf{g}^T|^{-1/2} \exp\left(-\frac{\{y - \hat{y}_{\mathrm{MP}}(\mathbf{x})\}^2}{2\sigma_{\mathcal{D}}^2(\mathbf{x})}\right), \end{aligned} \tag{A7}$$

where now the variance of posterior predictive distribution of y is:

$$\sigma^2_{\mathcal{D}}(\mathbf{x}) = \frac{1}{\beta - \beta^2 \mathbf{g}^T(\mathbf{A} + \beta \mathbf{g}\mathbf{g}^T)^{-1}\mathbf{g}} = \frac{1}{\beta} + \mathbf{g}(\mathbf{x})^T \mathbf{A}^{-1}\mathbf{g}(\mathbf{x}), \tag{A8}$$

where we have now explicitly stated the dependency of $\sigma^2_{\mathcal{D}}(\mathbf{x})$ on $\mathbf{x}$. The right side of Equation (A7) follows straightforwardly using known results on multidimensional Gaussian integrals. In addition, the right side of Equation (A8) results conveniently via algebraic manipulation. Detailed results for Equations (A7) and (A8) can be found from the appendix of this manuscript. By now simply discarding the factor $(2\pi)^{q/2}|\mathbf{A} + \beta \mathbf{g}\mathbf{g}^T|^{-1/2}$ from the right side of Equation (A7) and adding a multiplying factor $\{2\pi\sigma^2_{\mathcal{D}}(\mathbf{x})\}^{-1/2}$, we can write $p(y|\mathbf{x}, \mathcal{D})$ as:

$$p(y|\mathbf{x}, \mathcal{D}) = \frac{1}{\sqrt{2\pi\sigma^2_{\mathcal{D}}(\mathbf{x})}} \exp\left(-\frac{\{y - \hat{y}_{\mathrm{MP}}(\mathbf{x})\}^2}{2\sigma^2_{\mathcal{D}}(\mathbf{x})}\right). \tag{A9}$$

One might have an issue with dropping out the non-constant factor $t(\mathbf{x}) \overset{\Delta}{=} (2\pi)^{q/2}|\mathbf{A} + \beta \mathbf{g}(\mathbf{x})\mathbf{g}(\mathbf{x})^T|^{-1/2}$ in Equation (A7) but this is not a problem, since it simply scales the distribution function of $y|\mathbf{x}, \mathcal{D}$ and the variance $\sigma^2_{\mathcal{D}}(\mathbf{x})$ is invariant to this effect. Regarding the Algorithm 2, the factor $t(\mathbf{x})$ is also irrelevant and does not affect the functionality of BMVI sampling. Thus, we have now that the conditional posterior predictive distribution of y, given an input datum $\mathbf{x}$ and data set $\mathcal{D}$ is $y|\mathbf{x}, \mathcal{D} \sim \mathcal{N}(\hat{y}_{\mathrm{MP}}(\mathbf{x}), \sigma^2_{\mathcal{D}}(\mathbf{x}))$.

In a special case, if we use a linear function as the prediction model, i.e., $f(\mathbf{x}; \boldsymbol{\theta}) = \mathbf{x}^T\boldsymbol{\theta} + w_0$, then it is easy to show that the Hessian matrix $\mathbf{A}$ of $S(\boldsymbol{\theta})$ in Equation (A5) has the form:

$$\mathbf{A} = \beta X^T X + \alpha \begin{bmatrix} 0 & \mathbf{0}_{1\times m} \\ \mathbf{0}_{m\times 1} & I_{m\times m} \end{bmatrix}, \tag{A10}$$

where $\mathbf{0}_{m\times 1}$ and $\mathbf{0}_{1\times m}$ are m-dimensional zero vectors and matrix X is defined as:

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1m} \\ 1 & x_{21} & x_{22} & \dots & x_{2m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nm} \end{pmatrix},$$

where i^{th} row contains the i^{th} input vector $\mathbf{x}_i$ (with term 1 corresponding to constant parameter θ_0). We see that $\mathbf{A}$ is a positive semidefinite matrix, implying the convexity of $S(\boldsymbol{\theta})$. This means the maximum posterior point $\boldsymbol{\theta}_{\mathrm{MP}}$ for a linear model is:

$$\boldsymbol{\theta}_{\mathrm{MP}} = \underset{\boldsymbol{\theta}\in\mathbb{R}^m, w_0\in\mathbb{R}}{\operatorname{argmin}} S(\boldsymbol{\theta}) = \left(X^T X + \frac{\alpha}{\beta}\begin{bmatrix} 0 & \mathbf{0}_{1\times m} \\ \mathbf{0}_{m\times 1} & I_{m\times m} \end{bmatrix}\right)^{-1} X^T \mathbf{y},$$

where $\mathbf{y}$ is a $n \times 1$ vector of output values. It follows that the variance of $p(y|\mathbf{x}, \mathcal{D})$ for a linear prediction model is:

$$\sigma^2_{\mathcal{D}}(\mathbf{x}) = \frac{1}{\beta} + \mathbf{x}^T \mathbf{A}^{-1}\mathbf{x}. \tag{A11}$$

Finally, we will present the derivations of the results in Equations (A7) and (A8). In the following equations, we denote $\mathbf{C} = \mathbf{A} + \beta \mathbf{g}\mathbf{g}^T$ and $D = \mathbf{g}^T\mathbf{C}^{-1}\mathbf{g}$. We will also take advantage of the following known results:

$$\int_{-\infty}^{\infty} \exp\left(-\frac{\lambda}{2}x^2\right) dx = \left(\frac{2\pi}{\lambda}\right)^{1/2},$$

$$\int_{\mathbb{R}^q} \exp\left(-\frac{1}{2}\boldsymbol{\theta}^T\mathbf{A}\boldsymbol{\theta} + \mathbf{h}^T\boldsymbol{\theta}\right) d\boldsymbol{\theta} = (2\pi)^{q/2}|\mathbf{A}|^{-1/2} \exp\left(\frac{1}{2}\mathbf{h}^T\mathbf{A}^{-1}\mathbf{h}\right),$$

where $\mathbf{A}$ is a real symmetric matrix, $\mathbf{h}$ and $\boldsymbol{\theta}$ are q-dimensional vectors, and the integration is over whole $\boldsymbol{\theta}$-space $\mathbb{R}^q$.

Equation (A7), closed form of $p(y|\mathbf{x}, \mathcal{D})$:

$$\begin{aligned}
p(y|\mathbf{x}, \mathcal{D}) &\propto \int_{\mathbb{R}^q} \exp\left(-\frac{\beta}{2}\left\{y - \hat{y}_{\mathrm{MP}}(\mathbf{x}) - \mathbf{g}^T\boldsymbol{\theta}\right\}^2 - \frac{1}{2}\boldsymbol{\theta}^T\mathbf{A}\boldsymbol{\theta}\right) d\boldsymbol{\theta} \\
&= \int_{\mathbb{R}^q} \exp\left(-\frac{\beta}{2}\{y - \hat{y}_{\mathrm{MP}}(\mathbf{x})\}^2 + \beta\{y - \hat{y}_{\mathrm{MP}}(\mathbf{x})\}\mathbf{g}^T\boldsymbol{\theta} - \frac{\beta}{2}\boldsymbol{\theta}^T\mathbf{g}\mathbf{g}^T\boldsymbol{\theta} - \frac{1}{2}\boldsymbol{\theta}^T\mathbf{A}\boldsymbol{\theta}\right) d\boldsymbol{\theta} \\
&= \exp\left(-\frac{\beta}{2}\{y - \hat{y}_{\mathrm{MP}}(\mathbf{x})\}^2\right) \int_{\mathbb{R}^q} \exp\left(-\frac{1}{2}\boldsymbol{\theta}^T\mathbf{C}\boldsymbol{\theta} + \beta\{y - \hat{y}_{\mathrm{MP}}(\mathbf{x})\}\mathbf{g}^T\boldsymbol{\theta}\right) d\boldsymbol{\theta} \\
&= \exp\left(-\frac{\beta}{2}\{y - \hat{y}_{\mathrm{MP}}(\mathbf{x})\}^2\right)\left[(2\pi)^{q/2}|\mathbf{C}|^{-1/2}\exp\left(\frac{1}{2}\beta\{y - \hat{y}_{\mathrm{MP}}(\mathbf{x})\}\mathbf{g}^T\mathbf{C}^{-1}\beta\{y - \hat{y}_{\mathrm{MP}}(\mathbf{x})\}\mathbf{g}\right)\right] \\
&= \exp\left(-\frac{\beta}{2}\{y - \hat{y}_{\mathrm{MP}}(\mathbf{x})\}^2\right)\left[(2\pi)^{q/2}|\mathbf{C}|^{-1/2}\exp\left(\frac{1}{2}\beta^2\{y - \hat{y}_{\mathrm{MP}}(\mathbf{x})\}^2 D\right)\right] \\
&= (2\pi)^{q/2}|\mathbf{C}|^{-1/2}\exp\left(-\frac{\beta}{2}\{y - \hat{y}_{\mathrm{MP}}(\mathbf{x})\}^2\right)\exp\left((-D\beta)\left(-\frac{\beta}{2}\right)\{y - \hat{y}_{\mathrm{MP}}(\mathbf{x}))^2\right) \\
&= (2\pi)^{q/2}|\mathbf{C}|^{-1/2}\exp\left((1 - D\beta)\left(-\frac{\beta}{2}\right)\{y - \hat{y}_{\mathrm{MP}}(\mathbf{x})\}^2\right) \\
&= (2\pi)^{q/2}|\mathbf{A} + \beta\mathbf{g}\mathbf{g}^T|^{-1/2}\exp\left(-\frac{\beta - \beta^2\mathbf{g}^T(\mathbf{A} + \beta\mathbf{g}\mathbf{g}^T)^{-1}\mathbf{g}}{2}\{y - \hat{y}_{\mathrm{MP}}(\mathbf{x})\}^2\right) \\
&= (2\pi)^{q/2}|\mathbf{A} + \beta\mathbf{g}\mathbf{g}^T|^{-1/2}\exp\left(-\frac{\{y - \hat{y}_{\mathrm{MP}}(\mathbf{x})\}^2}{2\sigma^2_{\mathcal{D}}(\mathbf{x})}\right),
\end{aligned}$$

where $\sigma^2_{\mathcal{D}}(\mathbf{x}) = \left(\beta - \beta^2\mathbf{g}^T(\mathbf{A} + \beta\mathbf{g}\mathbf{g}^T)^{-1}\mathbf{g}\right)^{-1}$ ■ Note that integrating $p(y|\mathbf{x}, \mathcal{D})$ with respect to y gives:

$$\begin{aligned}
\int_{-\infty}^{\infty} p(y|\mathbf{x}, \mathcal{D})\, dy &\propto \int_{-\infty}^{\infty} (2\pi)^{q/2}|\mathbf{A} + \beta\mathbf{g}\mathbf{g}^T|^{-1/2}\exp\left(-\frac{\{y - \hat{y}_{\mathrm{MP}}(\mathbf{x})\}^2}{2\sigma^2_{\mathcal{D}}(\mathbf{x})}\right) dy \\
&= (2\pi)^{q/2}|\mathbf{A} + \beta\mathbf{g}\mathbf{g}^T|^{-1/2}\int_{-\infty}^{\infty}\exp\left(-\frac{\{y - \hat{y}_{\mathrm{MP}}(\mathbf{x})\}^2}{2\sigma^2_{\mathcal{D}}(\mathbf{x})}\right) dy \\
&= (2\pi)^{q/2}|\mathbf{A} + \beta\mathbf{g}\mathbf{g}^T|^{-1/2}\sqrt{2\pi\sigma^2_{\mathcal{D}}(\mathbf{x})},
\end{aligned}$$

which contains the reciprocal of $\{2\pi\sigma^2_{\mathcal{D}}(\mathbf{x})\}^{-1/2}$ we added in Equation (A9).

Equation (A8), closed form of $\sigma^2_{\mathcal{D}}(\mathbf{x})$:

$$\begin{aligned}
\sigma^2_{\mathcal{D}}(\mathbf{x}) &= \frac{1}{\beta - \beta^2\mathbf{g}^T(\mathbf{A} + \beta\mathbf{g}\mathbf{g}^T)^{-1}\mathbf{g}} \\
&= \frac{1}{\beta - \beta^2\mathbf{g}^T(\mathbf{A} + \beta\mathbf{g}\mathbf{g}^T)^{-1}\mathbf{g}} \times \frac{\mathbf{g}^T(\mathbf{I} + \beta\mathbf{A}^{-1}\mathbf{g}\mathbf{g}^T)\mathbf{g}}{\mathbf{g}^T(\mathbf{I} + \beta\mathbf{A}^{-1}\mathbf{g}\mathbf{g}^T)\mathbf{g}} \\
&= \frac{\mathbf{g}^T\mathbf{g} + \mathbf{g}^T\beta\mathbf{A}^{-1}\mathbf{g}\mathbf{g}^T\mathbf{g}}{\beta\mathbf{g}^T\mathbf{g} + \beta^2\mathbf{g}^T\mathbf{A}^{-1}\mathbf{g}\mathbf{g}^T\mathbf{g} - \beta^2\mathbf{g}^T(\mathbf{A} + \beta\mathbf{g}\mathbf{g}^T)^{-1}\mathbf{g}\mathbf{g}^T\mathbf{g} - \beta^2\mathbf{g}^T(\mathbf{A} + \beta\mathbf{g}\mathbf{g}^T)^{-1}\mathbf{g}\mathbf{g}^T(\beta\mathbf{A}^{-1}\mathbf{g}\mathbf{g}^T)\mathbf{g}} \\
&= \frac{1 + \mathbf{g}^T\beta\mathbf{A}^{-1}\mathbf{g}}{\beta + \beta^2\mathbf{g}^T\mathbf{A}^{-1}\mathbf{g} - \beta^2\mathbf{g}^T(\mathbf{A} + \beta\mathbf{g}\mathbf{g}^T)^{-1}\mathbf{g} - \beta^2\mathbf{g}^T(\mathbf{A} + \beta\mathbf{g}\mathbf{g}^T)^{-1}\mathbf{g}\mathbf{g}^T\beta\mathbf{A}^{-1}\mathbf{g}} \\
&= \frac{\frac{1}{\beta} + \mathbf{g}^T\mathbf{A}^{-1}\mathbf{g}}{1 + \beta\mathbf{g}^T\mathbf{A}^{-1}\mathbf{g} - \beta\mathbf{g}^T(\mathbf{A} + \beta\mathbf{g}\mathbf{g}^T)^{-1}\mathbf{g} - \beta\mathbf{g}^T(\mathbf{A} + \beta\mathbf{g}\mathbf{g}^T)^{-1}\mathbf{g}\mathbf{g}^T\beta\mathbf{A}^{-1}\mathbf{g}}.
\end{aligned}$$

Now in order for the result in Equation (A8) to hold, the denominator excluding term 1 in the above fraction should equate to 0:

$$\beta \mathbf{g}^T \mathbf{A}^{-1} \mathbf{g} - \beta \mathbf{g}^T (\mathbf{A} + \beta \mathbf{g}\mathbf{g}^T)^{-1} \mathbf{g} - \beta \mathbf{g}^T (\mathbf{A} + \beta \mathbf{g}\mathbf{g}^T)^{-1} \mathbf{g}\mathbf{g}^T \beta \mathbf{A}^{-1} \mathbf{g} = 0$$
$$\mathbf{g}^T \mathbf{A}^{-1} \mathbf{g} - \mathbf{g}^T (\mathbf{A} + \beta \mathbf{g}\mathbf{g}^T)^{-1} \mathbf{g} - \mathbf{g}^T (\mathbf{A} + \beta \mathbf{g}\mathbf{g}^T)^{-1} \mathbf{g}\mathbf{g}^T \beta \mathbf{A}^{-1} \mathbf{g} = 0$$

$$\mathbf{g}^T \mathbf{A}^{-1} \mathbf{g} - \mathbf{g}^T (\mathbf{A} + \beta \mathbf{g}\mathbf{g}^T)^{-1} \mathbf{g} = \mathbf{g}^T (\mathbf{A} + \beta \mathbf{g}\mathbf{g}^T)^{-1} \mathbf{g}\mathbf{g}^T \beta \mathbf{A}^{-1} \mathbf{g}$$
$$\mathbf{g}^T \mathbf{A}^{-1} \mathbf{g} - \mathbf{g}^T \mathbf{C}^{-1} \mathbf{g} = \beta \mathbf{g}^T \mathbf{C}^{-1} \mathbf{g}\mathbf{g}^T \mathbf{A}^{-1} \mathbf{g}$$
$$\mathbf{g}^T \mathbf{A}^{-1} \mathbf{g}\mathbf{g}^T (\mathbf{g}\mathbf{g}^T)^{-1} \mathbf{C} - \mathbf{g}^T \mathbf{C}^{-1} \mathbf{g}\mathbf{g}^T (\mathbf{g}\mathbf{g}^T)^{-1} \mathbf{C} = \beta \mathbf{g}^T \mathbf{C}^{-1} \mathbf{g}\mathbf{g}^T \mathbf{A}^{-1} \mathbf{g}\mathbf{g}^T (\mathbf{g}\mathbf{g}^T)^{-1} \mathbf{C}$$
$$\mathbf{g}^T \mathbf{A}^{-1} \mathbf{C} - \mathbf{g}^T = \beta \mathbf{g}^T \mathbf{C}^{-1} \mathbf{g}\mathbf{g}^T \mathbf{A}^{-1} \mathbf{C}$$
$$\mathbf{C}(\mathbf{g}\mathbf{g}^T)^{-1} \mathbf{g}\mathbf{g}^T \mathbf{A}^{-1} \mathbf{C} - \mathbf{C}(\mathbf{g}\mathbf{g}^T)^{-1} \mathbf{g}\mathbf{g}^T = \beta \mathbf{C}(\mathbf{g}\mathbf{g}^T)^{-1} \mathbf{g}\mathbf{g}^T \mathbf{C}^{-1} \mathbf{g}\mathbf{g}^T \mathbf{A}^{-1} \mathbf{C}$$
$$\mathbf{C}\mathbf{A}^{-1} \mathbf{C} - \mathbf{C} = \beta \mathbf{g}\mathbf{g}^T \mathbf{A}^{-1} \mathbf{C}$$
$$(\mathbf{A} + \beta \mathbf{g}\mathbf{g}^T) \mathbf{A}^{-1} (\mathbf{A} + \beta \mathbf{g}\mathbf{g}^T) - (\mathbf{A} + \beta \mathbf{g}\mathbf{g}^T) = \beta \mathbf{g}\mathbf{g}^T \mathbf{A}^{-1} (\mathbf{A} + \beta \mathbf{g}\mathbf{g}^T)$$
$$(\mathbf{A} + \beta \mathbf{g}\mathbf{g}^T)(\mathbf{I} + \beta \mathbf{A}^{-1} \mathbf{g}\mathbf{g}^T) - (\mathbf{A} + \beta \mathbf{g}\mathbf{g}^T) = \beta(\mathbf{g}\mathbf{g}^T + \beta \mathbf{g}\mathbf{g}^T \mathbf{A}^{-1} \mathbf{g}\mathbf{g}^T)$$
$$(\mathbf{A} + \beta \mathbf{g}\mathbf{g}^T)(\beta \mathbf{A}^{-1} \mathbf{g}\mathbf{g}^T) = \beta(\mathbf{g}\mathbf{g}^T + \beta \mathbf{g}\mathbf{g}^T \mathbf{A}^{-1} \mathbf{g}\mathbf{g}^T)$$
$$\beta(\mathbf{g}\mathbf{g}^T + \beta \mathbf{g}\mathbf{g}^T \mathbf{A}^{-1} \mathbf{g}\mathbf{g}^T) = \beta(\mathbf{g}\mathbf{g}^T + \beta \mathbf{g}\mathbf{g}^T \mathbf{A}^{-1} \mathbf{g}\mathbf{g}^T).$$

In other words,

$$\sigma_{\mathcal{D}}^2(\mathbf{x}) = \frac{1}{\beta - \beta^2 \mathbf{g}^T (\mathbf{A} + \beta \mathbf{g}\mathbf{g}^T)^{-1} \mathbf{g}} = \frac{1}{\beta} + \mathbf{g}^T \mathbf{A}^{-1} \mathbf{g} \quad \blacksquare$$

References

1. Särndal, C.E.; Swensson, B.; Wretman, J. *Model Assisted Survey Sampling (Springer Series in Statistics)*; Springer: New York, NY, USA, 1992.
2. Fuller, W.A. *Sampling Statistics*, 1st ed.; John Wiley & Sons, Inc.: Hoboken, United States, 2009; pp. 16–29. [CrossRef]
3. Kangas, A.; Maltamo, M. *Forest Inventory Methodology and Applications*, 1st ed.; Springer: Dordrecht, The Netherlands, 2006. [CrossRef]
4. Cochran, W.G. *Sampling Techniques*, 3rd ed.; John Wiley: Hoboken, NJ, USA, 1977.
5. Loetsch, F.; Haller, K.E. *Forest Inventory Vol. 1, Statistics of Forest Inventory and Information from Aerial Photographs*; BLV Verlagsgesellschaft: Munich, Germany, 1964.
6. Kondo, M.C.; Bream, K.D.; Barg, F.K.; Branas, C.C. A random spatial sampling method in a rural developing nation. *BMC Public Health* **2014**, *14*, 338. [CrossRef] [PubMed]
7. Pennanen, O.; Mäkelä, O. *Raakapuukuljetusten Kelirikkohaittojen Vähentäminen, Metsätehon Raportti*; Technical Report 153; Metsäteho Ltd.: Vantaa, Finland, 2003.
8. Pohjankukka, J.; Nevalainen, P.; Pahikkala, T.; Hyvönen, E.; Sutinen, R.; Hänninen, P.; Heikkonen, J. Arctic soil hydraulic conductivity and soil type recognition based on aerial gamma-ray spectroscopy and topographical data. In Proceedings of the 22nd International Conference on Pattern Recognition (ICPR 2014), Stockholm, Sweden, 24–28 August 2014; Borga, M., Heyden, A., Laurendeau, D., Felsberg, M., Boyer, K., Eds.; pp. 1822–1827. [CrossRef]
9. Pohjankukka, J.; Nevalainen, P.; Pahikkala, T.; Hyvönen, E.; Middleton, M.; Hänninen, P.; Ala-Ilomäki, J.; Heikkonen, J. Predicting Water Permeability of the Soil Based on Open Data. In Proceedings of the 10th International Conference on Artificial Intelligence Applications and Innovations (AIAI 2014), Rhodes, Greece, 19–21 September 2014; Lazaros, I., Ilias, M., Harris, P., Eds.; IFIP Advances in Information and Communication Technology; Springer: Berlin/Heidelberg, Germany, 2014; Volume 436, pp. 436–446. [CrossRef]
10. Pohjankukka, J.; Riihimäki, H.; Nevalainen, P.; Pahikkala, T.; Ala-Ilomäki, J.; Hyvönen, E.; Varjo, J.; Heikkonen, J. Predictability of Boreal Forest Soil Bearing Capacity by Machine Learning. *J. Terramech.* **2016**, *68*, 1–8. [CrossRef]
11. Tomppo, E.; Katila, M.; Mäkisara, K.; Peräsaari, J. *Multi-Source National Forest Inventory—Methods and Applications*; Managing Forest Ecosystems; Springer: Berlin, Germany, 2008; Volume 18. [CrossRef]
12. Wallner, A.; Elatawneh, A.; Schneider, T.; Kindu, M.; Ossig, B.; Knoke, T. Remotely sensed data controlled forest inventory concept. *Eur. J. Remote. Sens.* **2018**, *51*, 75–87. [CrossRef]
13. McRoberts, R.E.; Tomppo, E.O. Remote sensing support for national forest inventories. ForestSAT Special Issue. *Remote. Sens. Environ.* **2007**, *110*, 412–419. [CrossRef]
14. Puliti, S.; Ene, L.T.; Gobakken, T.; Næsset, E. Use of partial-coverage UAV data in sampling for large scale forest inventories. *Remote. Sens. Environ.* **2017**, *194*, 115–126. [CrossRef]

15. Abegg, M.; Kükenbrink, D.; Zell, J.; Schaepman, M.E.; Morsdorf, F. Terrestrial Laser Scanning for Forest Inventories—Tree Diameter Distribution and Scanner Location Impact on Occlusion. *Forests* **2017**, *8*, 184. [CrossRef]
16. Kangas, A.; Astrup, R.; Breidenbach, J.; Fridman, J.; Gobakken, T.; Korhonen, K.T.; Maltamo, M.; Nilsson, M.; Nord-Larsen, T.; Næsset, E.; et al. Remote sensing and forest inventories in Nordic countries–roadmap for the future. *Scand. J. For. Res.* **2018**, *33*, 397–412. [CrossRef]
17. White, J.; Coops, N.; Wulder, M.; Vastaranta, M.; Hilker, T.; Tompalski, P. Remote Sensing Technologies for Enhancing Forest Inventories: A Review. *Can. J. Remote. Sens.* **2016**, *42*, 619–641. [CrossRef]
18. Saukkola, A.; Melkas, T.; Riekki, K.; Sirparanta, S.; Peuhkurinen, J.; Holopainen, M.; Hyyppä, J.; Vastaranta, M. Predicting Forest Inventory Attributes Using Airborne Laser Scanning, Aerial Imagery, and Harvester Data. *Remote Sens.* **2019**, *11*, 797. [CrossRef]
19. Saad, R.; Wallerman, J.; Holmgren, J.; Lämås, T. Local pivotal method sampling design combined with micro stands utilizing airborne laser scanning data in a long term forest management planning setting. *Silva Fenn.* **2016**, *50*, 1414. [CrossRef]
20. Grafström, A.; Lundström, N.L.; Schelin, L. Spatially Balanced Sampling through the Pivotal Method. *Biometrics* **2012**, *68*, 514–520. [CrossRef] [PubMed]
21. Grafström, A.; Ringvall, A.H. Improving forest field inventories by using remote sensing data in novel sampling designs. *Can. J. For. Res.* **2013**, *43*, 1015–1022. [CrossRef]
22. Grafström, A.; Schelin, L. How to Select Representative Samples. *Scand. J. Stat.* **2014**, *41*, 277–290. [CrossRef]
23. Grafström, A.; Zhao, X.; Nylander, M.; Petersson, H. A new sampling strategy for forest inventories applied to the temporary clusters of the Swedish national forest inventory. *Can. J. For. Res.* **2017**, *47*, 1161–1167. [CrossRef]
24. Räty, M.; Heikkinen, J.; Kangas, A. Assessment of sampling strategies utilizing auxiliary information in large-scale forest inventory. *Can. J. For. Res.* **2018**, *48*, 749–757. [CrossRef]
25. Räty, M.; Kangas, A.S. Effect of permanent plots on the relative efficiency of spatially balanced sampling in a national forest inventory. *Ann. For. Sci.* **2019**, *76*, 20. [CrossRef]
26. Katila, M.; Heikkinen, J. Reducing error in small-area estimates of multi-source forest inventory by multi-temporal data fusion. *For. Int. J. For. Res.* **2020**, *93*, 471–480. [CrossRef]
27. Ruotsalainen, R.; Pukkala, T.; Kangas, A.; Vauhkonen, J.; Tuominen, S.; Packalen, P. The effects of sample plot selection strategy and the number of sample plots on inoptimality losses in forest management planning based on airborne laser scanning data. *Can. J. For. Res.* **2019**, *49*, 1135–1146. [CrossRef]
28. Räty, M.; Kuronen, M.; Myllymäki, M.; Kangas, A.; Mäkisara, K.; Heikkinen, J. Comparison of the local pivotal method and systematic sampling for national forest inventories. *For. Ecosyst.* **2020**, *7*, 54. [CrossRef]
29. de Gruijter, J.; Brus, D.; Bierkens, M.; Knotters, M. *Sampling for Natural Resource Monitoring*; Springer: Berlin, Germany, 2006. [CrossRef]
30. Brus, D. *Spatial Sampling with R*; Chapman and Hall/CRC: Boca Raton, FL, USA, 2022. [CrossRef]
31. Rasmussen, C.E.; Williams, C.K.I. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*; The MIT Press: Cambridge, MA, USA, 2005.
32. Garnett, R.; Osborne, M.A.; Roberts, S.J. Bayesian Optimization for Sensor Set Selection. In Proceedings of the 9th ACM/IEEE International Conference on Information Processing in Sensor Networks, Stockholm, Sweden, 12–16 April 2010; Association for Computing Machinery: New York, NY, USA, 2010; IPSN '10, pp. 209–219. [CrossRef]
33. Flynn, E.B.; Todd, M.D. A Bayesian approach to optimal sensor placement for structural health monitoring with application to active sensing. *Mech. Syst. Signal Process.* **2010**, *24*, 891–903. [CrossRef]
34. Liu, B. An Introduction to Bayesian Techniques for Sensor Networks. In Proceedings of the Wireless Algorithms, Systems, and Applications, Beijing, China, 15–17 August 2010; Pandurangan, G., Anil Kumar, V.S., Ming, G., Liu, Y., Li, Y., Eds.; Springer: Berlin/Heidelberg, Germany, 2010; pp. 307–313.
35. Tomppo, E.; Heikkinen, J.; Henttonen, H.; Ihalainen, A.; Katila, M.; Mäkelä, H.; Tuomainen, T.; Vainikainen, N. *Designing and Conducting a Forest Inventory—Case: 9th National Forest Inventory of Finland*, 1st ed.; Managing Forest Ecosystems 21; Springer: Dordrecht, The Netherlands, 2011. [CrossRef]
36. Metsäntutkimuslaitos. *Valtakunnan Metsien 11. Inventoinnin Maastotyöohje*; Metla: Joensuu, Finland, 2009.
37. Haara, A.; Kangas, A.; Tuominen, S. Economic losses caused by tree species proportions and site type errors in forest management planning. *Silva Fenn.* **2019**, *53*, 10089. [CrossRef]
38. Næsset, E. Accuracy of forest inventory using airborne laser scanning: Evaluating the first nordic full-scale operational project. *Scand. J. For. Res.* **2004**, *19*, 554–557. [CrossRef]
39. Packalén, P.; Maltamo, M. Predicting the Plot Volume by Tree Species Using Airborne Laser Scanning and Aerial Photographs. *For. Sci.* **2006**, *52*, 611–622. [CrossRef]
40. Packalén, P.; Maltamo, M. Estimation of species-specific diameter distributions using airborne laser scanning and aerial photographs. *Can. J. For. Res.* **2008**, *38*, 1750–1760. [CrossRef]
41. Yengoh, G.T.; Dent, D.; Olsson, L.; Tengberg, A.E.; Tucker, C.J. *Use of the Normalized Difference Vegetation Index (NDVI) to Assess Land Degradation at Multiple Scales: Current Status, Future Trends, and Practical Considerations*, 1st ed.; Springer Publishing Company, Incorporated: New York, NY, USA, 2015. [CrossRef]
42. Haralick, R.M.; Shanmugam, K.; Dinstein, I. Textural Features for Image Classification. *Syst. Man Cybern. IEEE Trans.* **1973**, *SMC-3*, 610–621. [CrossRef]

43. Pohjankukka, J. Machine Learning Approaches for Natural Resource Data. Ph.D. Thesis, University of Turku, Turku, Finland, 2018.
44. Pohjankukka, J.; Tuominen, S.; Pitkänen, J.; Pahikkala, T.; Heikkonen, J. Comparison of estimators and feature selection procedures in forest inventory based on airborne laser scanning and digital aerial imagery. *Scand. J. For. Res.* **2018**, *33*, 681–694. [CrossRef]
45. Racine, E.; Coops, N.; Bégin, J. Tree species, crown cover, and age as determinants of the vertical distribution of airborne LiDAR returns. *Trees* **2021**, *35*, 1845–1861. [CrossRef]
46. Kansanen, K.; Packalen, P.; Lähivaara, T.; Seppänen, A.; Vauhkonen, J.; Maltamo, M.; Mehtätalo, L. Refining and evaluating a horvitz-thompson-like stand density estimator in individual tree detection based on airborne laser scanning. *Can. J. For. Res.* **2022**, *52*, 527–538. [CrossRef]
47. Beland, M.; Parker, G.; Sparrow, B. On promoting the use of lidar systems in forest ecosystem research. *For. Ecol. Manag.* **2019**, *450*, 117484. [CrossRef]
48. Tuominen, S.; Balazs, A.; Kangas, A. Comparison of photogrammetric canopy models from archived and made-to-order aerial imagery in forest inventory. *Silva Fenn.* **2020**, *54*, 10291. [CrossRef]
49. Pahikkala, T.; Airola, A.; Salakoski, T. Speeding Up Greedy Forward Selection for Regularized Least-Squares. In Proceedings of the 2010 Ninth International Conference on Machine Learning and Applications, Washington, DC, USA, 12–14 December 2010; pp. 325–330. [CrossRef]
50. Deville, J.C.; Tillé, Y. Unequal probability sampling without replacement through a splitting method. *Biometrika* **1998**, *85*, 89–101. [CrossRef]
51. Bishop, C.M. *Neural Networks for Pattern Recognition*; Oxford University Press: Oxford, UK, 1995.
52. MacKay, D.J.C. Information-Based Objective Functions for Active Data Selection. *Neural Comput.* **1992**, *4*, 590–604. [CrossRef]
53. MacKay, D.J.C. The Evidence Framework Applied to Classification Networks. *Neural Comput.* **1992**, *4*, 720–736. [CrossRef]
54. MacKay, D.J.C. Bayesian Interpolation. *Neural Comput.* **1992**, *4*, 415–447. [CrossRef]
55. Neal, R.M. *Bayesian Learning for Neural Networks*; Springer: Berlin/Heidelberg, Germany, 1996. [CrossRef]
56. Xia, G.; Miranda, M.L.; Gelfand, A.E. Approximately optimal spatial design approaches for environmental health data. *Environmetrics* **2006**, *17*, 363–385. [CrossRef]
57. Chipeta, M.; Terlouw, D.; Phiri, K.; Diggle, P. Inhibitory geostatistical designs for spatial prediction taking account of uncertain covariance structure. *Environmetrics* **2017**, *28*, e2425. [CrossRef]
58. Müller, W.G. *Collecting Spatial Data: Optimum Design of Experiments for Random Fields*, 3rd ed.; Springer: Berlin/Heidelberg, Germany, 2007. [CrossRef]
59. Zhu, Z.; Stein, M.L. Spatial sampling design for prediction with estimated parameters. *J. Agric. Biol. Environ. Stat.* **2006**, *11*, 24–44. [CrossRef]
60. Diggle, P.; Lophaven, S. Bayesian Geostatistical Design. *Scand. J. Stat.* **2006**, *33*, 53–64. [CrossRef]
61. Snoek, J.; Larochelle, H.; Adams, R.P. Practical Bayesian Optimization of Machine Learning Algorithms. In Proceedings of the 25th International Conference on Neural Information Processing Systems—Volume 2, Lake Tahoe, NV, USA, 3–6 December 2012; Curran Associates Inc.: Red Hook, NY, USA, 2012; NIPS'12, pp. 2951–2959.
62. Osborne, M.A. Bayesian Gaussian Processes for Sequential Prediction, Optimisation and Quadrature. Ph.D. Thesis, Oxford University, Oxford, UK, 2010.
63. Werner, J.; Müller, G. *Spatio-Temporal Design*; John Wiley & Sons, Ltd.: Hoboken, United States, 2012.
64. BMVI. Bayesian Maximum Variance Inclusion—Python Implementation. 2019. Available online: https://github.com/jjepsuomi/Bayesian-maximum-variance-inclusion (accessed on 23 September 2019).
65. Bishop, C.M. *Pattern Recognition and Machine Learning (Information Science and Statistics)*; Springer: Secaucus, NJ, USA, 2006.
66. Vapnik, V.N. *Statistical Learning Theory*; Wiley-Interscience: Hoboken, NJ, USA, 1998; Volume 1.
67. Murphy, K.P. *Machine Learning: A Probabilistic Perspective*; The MIT Press: Cambridge, MA, USA, 2012.
68. Gelman, A.; Carlin, J.B.; Stern, H.S.; Dunson, D.B.; Vehtari, A.; Rubin, D.B. *Bayesian Data Analysis*, 3rd ed.; Chapman & Hall/CRC Texts in Statistical Science, Taylor & Francis: Boca Raton, FL, USA, 2013.
69. Nabney, I.T. *NETLAB: Algorithms for Pattern Recognition*; Springer: London, UK, 2004.
70. Bazaraa, M.S. *Nonlinear Programming: Theory and Algorithms*, 3rd ed.; Wiley Publishing: Hoboken, NJ, USA, 2013. [CrossRef]