Bayesian Evaluation of Smartphone Applications for Forest Inventories in Small Forest Holdings

Andrej Ficko

Department of Forestry and Renewable Forest Resources, Biotechnical Faculty, University of Ljubljana, Večna pot 83, 1000 Ljubljana, Slovenia; andrej.ficko@bf.uni-lj.si; Tel.: +386-13-20-3502

Received: 9 October 2020; Accepted: 27 October 2020; Published: 29 October 2020

Abstract: There are increasingly advanced mobile applications for forest inventories on the market. Small enterprises and nonprofessionals may find it difficult to opt for a more sophisticated application without comparing it to an established standard. In a small private forest holding (19 ha, 4 stands, 61 standing points), we compared TRESTIMA, a computer vision-based mobile application for stand inventories, to MOTI, a smartphone-based relascope, in measuring the number of stems (N) and stand basal area (G). Using a Bayesian approach, we (1) weighted evidence for the hypothesis of no difference in N and G between TRESTIMA and MOTI relative to the hypothesis of difference, and (2) weighted evidence for the hypothesis of overestimating versus underestimating N and G when using TRESTIMA compared to MOTI. The results of the Bayesian tests were then compared to the results of frequentist tests after the p-values of paired sample t-tests were calibrated to make both approaches comparable. TRESTIMA consistently returned higher N and G, with a mean difference of +305.8 stems/ha and +5.8 m²/ha. However, Bayes factors (BF₁₀) suggest there is only moderate evidence for the difference in N (BF₁₀ = 4.061) and anecdotal evidence for the difference in G (BF₁₀ = 1.372). The frequentist tests returned inconclusive results, with p-values ranging from 0.03 to 0.13. After calibration of the p-values, the frequentist tests suggested rather small odds for the differences between the applications. Conversely, the odds of overestimating versus underestimating N and G were extremely high for TRESTIMA compared to MOTI. In a small forest holding, Bayesian evaluation of differences in stand parameters can be more helpful than frequentist analysis, as Bayesian statistics do not rely on asymptotics and can answer more specific hypotheses.

Keywords: forest inventory; mobile applications; relascope; private forest owners; Bayesian informative hypotheses; accuracy; forest management planning

1. Introduction

In measuring forest stands, mobile applications have proven to be a valuable alternative to traditional field measurement devices. The number of mobile applications for forest inventories has increased significantly in the last few years, and information technology development has enabled new techniques such as virtual reality and augmented reality to be implemented in hand-held devices [1–5]. Some of the advantages of mobile applications are user-friendliness and fast processing of the collected data, which may be particularly useful for inexperienced surveyors such as private forest owners. Although private forest owners are not always interested in wood production [6], they manage a sizeable part of private forests in Europe and the US, and they contribute a significant share to the supply of round wood in these countries (see, e.g., [7]). Proper estimation of growing stock, stand density, and tree species composition are prerequisites for sustainable forest management and the sustainable supply of wood and other ecosystem services.

One of the mobile applications for stand inventories that is attractive to inexperienced users because of its simplicity is the TRESTIMA mobile application [8]. The advantage of TRESTIMA is that the user
can estimate stand parameters just by shooting photos of the desired forest area. The photos are then sent to a cloud service for automated image analysis. Tree trunks and tree species are automatically recognized using computer vision, and stand parameters such as basal area, stem count, volume, median diameter and stand height are calculated using predefined site- and species-specific height and volume functions. In measuring stand basal area, the application uses the proportional-to-size sampling principle of the Bitterlich relascope \[9\] with a dynamic basal area factor, where a single tree represents 0.6 to 1.4 \(m^2/ha\). Similar to the Spiegel Relaskop, stand basal area is calculated by multiplying the number of trees by the corresponding basal area factor for each tree. Stand volume is calculated by the default volume functions based on the measured median tree height or the corresponding height for each tree species from previously measured similar forests stored in the cloud. The accuracy of TRESTIMA has been tested for different tree species compositions and stand structures in Europe and Russia and was found to be similar to that of traditional angle counting methods \[10–12\]. The TRESTIMA forest inventory system has mainly been used by entrepreneurs and forest companies, but less so by private and family forest owners \[8\].

Among traditional dendrometers, the Spiegel Relaskop and various angle gauges have been widely utilized. The principle of the Spiegel Relaskop has been implemented in many mobile applications \[13\]. MOTI, for example, is an application that combines a digital relascope with stem counting and tree height measuring capabilities \[14,15\]. It provides estimates of stand basal area that are “as good as, if not better than, the analog Bitterlich method”, with automatic correction for slope (see \[16\] p. vii). MOTI cannot reach the level of precision of a Vertex clinometer for tree height unless far enough from the tree \[16\]. It has proven to be a valuable replacement for the Spiegel Relaskop in rapid stand parameter estimations, with excellent accuracy in stands similar to ours (e.g., \[17,18\]). MOTI has been considered as a standard private forest stand inventory tool by several private forest enterprises in Switzerland \[19\], providing inexpensive stand parameter estimates needed for further simulations \[20\].

When a private forest owner is deciding on a new mobile application for a forest inventory, he expects the new application to be better than the previous one in at least one criterion, without losing accuracy. Accuracy in the eyes of forest owners may be expressed relative to the accuracy of the second-best option available and not in the absolute sense of the allowable tolerance of the sampling error. This means that an alternative application that does not return excessively deviating estimates but is easier, faster, or better in any other aspect may be the preferred application. In more statistical language, the interest of the layman is not to test whether the estimates obtained by the new application deviate significantly from the “true” population value (i.e., whether the null hypothesis of no differences between the “true” population value of the parameter and the value obtained by the application can be rejected) but which one of the two possible assertions about the mean value of the parameter under investigation is more likely, i.e., the means of the stand parameters from two competing applications are equal (H\(_0\)) or the means are not equal (H\(_1\)). Stated differently, in comparing two applications, forest owners do not treat the stand parameter under estimation as fixed with unknown bounds, depicted by the confidence interval as in the frequentist analysis, but rather as an unknown value that is highly likely within a credible interval in the Bayesian sense. This research question calls for a Bayesian approach to testing the differences between two alternative applications. In Bayesian hypothesis testing, the probability of the null (H\(_0\)) and the alternative hypothesis (H\(_1\)), based on the observed data, can be estimated (e.g., \[21\]), while in frequentist testing, the null hypothesis can only be rejected and a conclusion reached that there is a statistically significant difference between the two methods.

There are also other advantages of Bayesian analysis when assessing a new application. In Bayesian hypothesis testing, the hypothesis of a difference between two applications does not need to be precise but can also be informative. This means that the researcher can specify the expected relations between parameters and may include effect sizes \[22\] (see \[23\] for details). In measuring stands, forest owners are not just interested in whether the data provide evidence in favor of H\(_1\) (be it one-sided or two-sided H\(_1\)), but they might also be interested in testing the plausibility of several alternative hypotheses.
For instance, Bayesian analysis can weigh the evidence for the hypothesis of one application returning higher means than the other application ($H_2$) versus the hypothesis of one application returning lower means than the other application ($H_3$). The dilemma is not trivial; forest owners are likely to be more concerned about overestimating versus underestimating stand parameters when using a new application than just determining whether a difference between the two applications exists. The difference can be statistically significant but small in size.

Another aspect that should be carefully taken into account when testing hypotheses is the sampling method and sample size. Forest stands are usually too large to obtain all the measurements. Instead, a subset of all the measurements in the stand is used, from which the characteristics of the stand are inferred [24]. In a forest inventory, probability sampling ensures that all sampling units (e.g., trees or standpoints) have known chances of being selected. The advantage of using randomization is the absence of systematic and sampling bias. If random selection is made properly, the sample is representative of the stand, and the sample mean is expected to be an unbiased estimate of the population mean. However, probability sampling and large samples require more skill and more funds compared to nonprobability sampling. Private forest owners are likely to use nonprobability sampling because they lack knowledge on how to correctly conduct random sampling or simply because of the intuitiveness of nonprobability sampling techniques, such as purposive sampling, subjective sampling, typical case sampling, convenience sampling and various other combinations. The ability to go out in the forest with a new application, take a few photos, and instantly obtain stand parameter estimates is attractive. Although it is clear that such an approach does not guarantee unbiased estimates of the population values and that Type II errors are more likely to occur when sample sizes are small and frequentist statistics is used, nonprobability sampling and small samples do not prevent private forest owners from asking the following two questions before purchasing a new application: (1) Does the new application provide stand parameters that are as credible as those estimated by traditional field measurement devices? (2) What is the probability of overestimating or underestimating stand parameters when using the new application?

Since both investigated applications use different standing volume calculation methods, and MOTI uses Swiss tariffs, the applications were compared only with respect to stand basal area and stem count per ha. Specifically, we used Bayesian hypothesis testing to (1) weight evidence for the hypothesis of no difference between the TRESTIMA image analysis application and field measurements with the MOTI digital relascope, relative to the hypothesis of difference; (2) weight evidence for the hypothesis of overestimating versus underestimating stand parameters when using TRESTIMA compared to MOTI, assuming that MOTI is an accurate replacement for the Spiegel Relaskop [16].

2. Materials and Methods

2.1. Study Area and Field Measurements

The study area encompasses 19.4 ha of private forests in two forest compartments in the Kočevje forest management region, Slovenia, approximately 34 km from Ljubljana (45.81° N, 14.67° E), and consists of four stands (Z045, Z046, Z048_29, and Z048_30; Figure 1). All stands were mixed Norway spruce, silver fir, European beech, and Scots pine stands in the timber phase, with Norway spruce dominating the tree species composition. Stands were selected based on the similarity of site conditions and the dominance of Norway spruce and Scots pine, which TRESTIMA recognizes accurately [10]. The number of sampling points was 61. The number of photos taken in each stand was based on the size of the stand, stand structure, and the recommendation that at least 10 photos are needed per stand [25] and was as follows [26]: 17 photos in stand Z045 (6.5 ha), 10 photos in Z046 (2.51 ha), 19 photos in Z048_29 (5.46 ha), and 15 photos in Z048_30 (4.93 ha). All photos were taken by circling the stands, close to their borders, and shooting photos towards the center of the stand. The sampling points from which the photos were taken were chosen subjectively so that they were approximately 30 to 50 m apart. The GPS on a mobile phone, printed stand maps, and the GPS
on a hand-held device were used to record the positions of the standpoints. After the stand was photographed, all photos were uploaded to the cloud. Since there were no specific tree species profiles available for Slovenia, we chose Germany as the most suitable one, in which TRESTIMA recognized Norway spruce and Scots pine, while all other species except for silver fir were classified as other species (Figure 2). The application was unable to discern between silver fir and treated silver fir as spruce. The application calculated tree height for Norway spruce by utilizing the data it has from previously measured similar forests. For Scots pine and other tree species, tree height was measured by a clinometer and manually entered into the application since TRESTIMA did not determine the corresponding height automatically. After one to two minutes, basal area, stem count, stand volume, median diameter, and diameter distribution were calculated, and a report with all parameters was received on the mobile phone.

Figure 1. Study area with the four study stands (shades of green) and 61 sampling points.

Approximately 10 m away from the TRESTIMA sampling points, in the direction the photographs were taken, sampling points for digital Bitterlich sampling with MOTI [14] were selected (Figure 3). The distance of 10 m was chosen because the mobile phone camera can only capture 60 to 70 degrees of a full circle, depending on the camera lens, while in conventional Bitterlich sampling, a full circle is surveyed. Moreover, TRESTIMA uses smaller basal area factors and, thus, compared to the relascope, picks up tree trunks that are farther away, which means that the sampling points for MOTI were chosen approximately in the middle of the circular sector photographed by TRESTIMA.
10 should be considered as anecdotal, mild, moderate, and strong evidence, respectively, in favor of \( H_0 \) \([27,28]\), although the thresholds are fully arbitrary and have no statistical meaning. We used proportion wheels to visualize the strength of evidence for \( H_0 \). In proportion wheels, Bayes factor BF \(_{01}\) indicates how likely the hypothesis that TRESTIMA equals MOTI (i.e., the null hypothesis \( H_0 \)) is in comparison to the hypothesis of a difference (i.e., the two-sided alternative hypothesis \( H_1 \)). For instance, if Bayes factor BF \(_{01}\) equals 10, \( H_0 \) is 10-times more likely than \( H_1 \), given the data. BF \(_{01}\) between 1 and 3, between 3 and 6, between 6 and 10, and above 10 should be considered as anecdotal, mild, moderate, and strong evidence, respectively, in favor of \( H_0 \) \([27,28]\), although the thresholds are fully arbitrary and have no statistical meaning. We used proportion wheels to visualize the strength of evidence for \( H_0 \). In proportion wheels, Bayes factor BF \(_{01}\) was transformed to a magnitude between 0 and 1 and plotted as the proportion of a circular area.

In setting the probabilities of \( H_0 \) and \( H_1 \) before observing the data (i.e., prior probabilities, or priors \( P(H_0) \) and \( P(H_1) \); Equation (1)), we used a default Cauchy prior width of 0.707, which means that we

2.2. **Bayesian and Frequentist Paired Sample t-Test**

To test the relative evidence in the stand data about whether stem number and stand basal area estimated by TRESTIMA and MOTI are equal or different, we used a Bayesian paired sample \( t \)-test. In the Bayesian paired sample \( t \)-test, the Bayes factor BF \(_{01}\) indicates how likely the hypothesis that TRESTIMA equals MOTI (i.e., the null hypothesis \( H_0 \)) is in comparison to the hypothesis of a difference (i.e., the two-sided alternative hypothesis \( H_1 \)). For instance, if Bayes factor BF \(_{01}\) equals 10, \( H_0 \) is 10-times more likely than \( H_1 \), given the data. BF \(_{01}\) between 1 and 3, between 3 and 6, between 6 and 10, and above 10 should be considered as anecdotal, mild, moderate, and strong evidence, respectively, in favor of \( H_0 \) \([27,28]\), although the thresholds are fully arbitrary and have no statistical meaning. We used proportion wheels to visualize the strength of evidence for \( H_0 \). In proportion wheels, Bayes factor BF \(_{01}\) was transformed to a magnitude between 0 and 1 and plotted as the proportion of a circular area.

In setting the probabilities of \( H_0 \) and \( H_1 \) before observing the data (i.e., prior probabilities, or priors \( P(H_0) \) and \( P(H_1) \); Equation (1)), we used a default Cauchy prior width of 0.707, which means that we
assume \( H_0 \) and \( H_1 \) are equally likely and that we were 50% certain that the effect of an application would be between \(-0.707\) and \(0.707\).

\[
\frac{P(H_0|\text{data})}{P(H_1|\text{data})} = BF_{01} \cdot \frac{P(H_0)}{P(H_1)}
\]  

(1)

Bayesian error probabilities \( P(H_0|\text{data}) \) and \( P(H_1|\text{data}) \) (also called posterior probabilities), which quantify the support for \( H_0 \) and \( H_1 \) after observing the data, are then calculated by multiplying Bayes factor \( BF_{01} \) with prior odds (Equation (1)). It is well known that the choice of priors affects the resulting Bayes factor. Therefore, the sensitivity of the Bayesian test was checked by changing the Cauchy prior width and inspecting \( BF_{01} \). In addition, we conducted sequential analysis, in which we inspected the strength of the evidence supporting the null hypothesis (\( BF_{01} \)) with respect to sample size and the range of plausible values for the prior.

The results of the Bayesian paired sample \( t \)-test were then compared to the results of the frequentist paired sample \( t \)-test. The paired sample \( t \)-test tests the null hypothesis (\( H_0 \)) that the population mean of the difference between paired observations equals zero. In more plain language, we tested whether the differences between paired observations converge to zero if measurements are repeated indefinitely.

In frequentist analysis, the \( p \)-value is the central point answering this question; the \( p \)-value is the probability of the test statistic being more extreme than or equal to the observed results of a statistical hypothesis test. However, the \( p \)-value has frequently been misinterpreted as the probability that \( H_0 \) is true or the probability that \( H_1 \) is false in a frequentist sense (e.g., [29]). To make the Bayesian and frequentist analyses comparable, we calibrated the \( p \)-value by calculating two measures: the Vovk–Sellke maximum \( p \)-ratio (VS–MPR) and the frequentist Type I error probability (\( \alpha(p) \)).

The VS–MPR is obtained by choosing shape \( \alpha \) of the \( p \)-value distribution under \( H_2 \), such that the obtained \( p \)-value is maximally diagnostic. The VS–MPR value is then the ratio of the densities at point \( p \) under \( H_0 \) and \( H_1 \) [29]. The bound \( 1/(e \cdot p \ln(p)) \), also called the Benjamin and Berger bound [30,31], is derived from the shape of the \( p \)-value distribution when \( p < 1/e \). Under \( H_0 \), the bound is uniform (0, 1), and under \( H_1 \), it is decreasing in \( p \), e.g., by a beta \((\alpha, 1)\) distribution, where \( 0 < \alpha < 1 \). For example, if the two-sided \( p \)-value equals 0.05, the Vovk–Sellke MPR equals 2.46, indicating that the \( p \)-value of 0.05 is, at most, 2.46 times more likely to occur under \( H_1 \) than under \( H_0 \). A statistical widget for calculating the VS–MPR was used [32] when the VS–MPR was not reported by the software.

The frequentist Type I error probability (\( \alpha(p) \)) for rejecting \( H_0 \) was calculated as \( \alpha(p) = (1 + (−e \cdot p \ln(p))^{−1})^{−1} \) [29]. This represents the probability of false rejection of \( H_0 \) at the two-sided \( p \)-value. For instance, if the two-sided \( p \)-value equals 0.05, the error probability in rejecting \( H_0 \) is 0.289. Alternatively, we could also say that in 28.9% of the \( t \)-tests, the \( p \)-value would be greater than 0.05.

2.3. Bayesian Informative Hypothesis Evaluation (Bain)

To weight evidence for the hypothesis of overestimating versus underestimating stand parameters when using TRESTIMA compared to MOTI, we used a Bayesian informative hypotheses evaluation (bain) paired samples \( t \)-test in which the following two alternative hypotheses were evaluated: \( H_2 \): mean_{TRESTIMA} > mean_{MOTI}; \( H_3 \): mean_{TRESTIMA} < mean_{MOTI}:

\[
\frac{P(H_2|\text{data})}{P(H_3|\text{data})} = BF_{23} \cdot \frac{P(H_2)}{P(H_3)}
\]  

(2)

\( BF_{23} \) is the ratio between how well the hypothesis of overestimation by TRESTIMA performs in explaining the data compared to the hypothesis of underestimation. \( P(H_2)/P(H_3) \) is the prior odds of these two hypotheses and \( P(H_2|\text{data})/P(H_3|\text{data}) \) is the posterior odds for these two hypotheses (Equation (2)). All tests were performed in JASP 0.11.1 [33], based on a sample of \( n = 4 \) stands and for \( n = 61 \) paired observations of basal area. The hypotheses can be summarized as follows (Table 1):
Forests 2020, 11, 1148

Table 1. Hypotheses about differences in the number of stems (N) and stand basal area (G) between TRESTIMA and MOTI.

| Hypotheses | Paired Sample t-Test | The Bain Paired Sample t-Test |
|-------------|----------------------|------------------------------|
| TRESTIMA = MOTI | H₀: N_TRESTIMA = N_MOTI | H₂: G_TRESTIMA > G_MOTI | H₀: N_TRESTIMA = G_MOTI | H₂: G_TRESTIMA > G_MOTI |
| Two-sided alternative hypothesis | H₁: N_TRESTIMA ≠ N_MOTI | H₁: G_TRESTIMA ≠ G_MOTI | Hypothesis that the population mean of the TRESTIMA estimates is greater than the population mean of the MOTI estimates. |
| The null hypothesis that the population mean of differences in N between the two applications equals 0. | The hypothesis that the population mean of differences in G between the two applications equals 0. | H₀: N_TRESTIMA < N_MOTI | H₂: G_TRESTIMA < G_MOTI |
| The hypothesis that the population mean of differences in N between the two applications does not equal 0. | The hypothesis that the population mean of differences in G between the two applications does not equal 0. | H₀: G_TRESTIMA < G_MOTI | H₂: G_TRESTIMA < G_MOTI |

Bayesian approach

BF₀₁ in Bayesian paired sample t-test

BF₀₂ in the Bayesian informative hypotheses evaluation (bain) paired sample t-test

p-value in paired sample t-test, the Vovk–Sellke maximum p–ratio (VS-MPR) and the Frequentist type I error probability (α(p))

N/A

p-value: The smaller the p-value, the smaller the probability of the t-statistic being more extreme than or equal to the observed t in the test.

N/A

VS-MPR: The bigger the VS-MPR, the more likely it is that the p-value occurs under H₁ than under H₀.

α(p): The bigger the α(p), the greater the probability of false rejection of H₀.

3. Results

3.1. Differences in the Number of Stems per Hectare between TRESTIMA and MOTI

TRESTIMA estimates the number of stems only at the stand level; therefore, a comparison of the number of stems was not possible at the standing point level. At the stand level, TRESTIMA returned, on average, a higher number of stems than MOTI. The mean difference in stand-wise paired comparisons was +305.8 trees/ha (Table 2).

Table 2. Descriptive statistics and normality tests for the number of stems and basal area estimated by TRESTIMA and MOTI based on stand-wise estimation (n = 4) and for all standing points (n = 61).

|            | N  | Mean | Median | SD  | SE  | p-Value of Shapiro–Wilk |
|------------|----|------|--------|-----|-----|-------------------------|
| TRESTIMA_N (ha) | 4  | 924.5| 947.5  | 130.4| 65.2| 0.77                    |
| MOTI_N (ha)    | 4  | 618.8| 593.5  | 112.5| 56.3| 0.42                    |
| TRESTIMA_N–MOTI_N (ha) | 4  | 305.8| 308.0  | 149.7| 74.8| 0.43                    |
| TRESTIMA_G (m²/ha) | 4  | 39.8 | 40.7   | 8.2  | 4.1 | 0.86                    |
| MOTI_G (m²/ha)  | 4  | 34.0 | 31.5   | 8.5  | 4.3 | 0.36                    |
| TRESTIMA_G–MOTI_G (m²/ha) | 4  | 5.8  | 4.0    | 5.0  | 2.5 | 0.19                    |
| TRESTIMA_G (m²/ha) | 61 | 41.2 | 41.2   | 15.3 | 2.0 | 0.37                    |
| MOTI_G (m²/ha)  | 61 | 35.5 | 34.0   | 11.7 | 1.5 | 0.00                    |
| TRESTIMA_G–MOTI_G (m²/ha) | 61 | 5.7  | 7.6    | 17.7 | 2.3 | 0.71                    |

The Bayesian paired sample t-test suggests that the data in the four stands moderately support the alternative hypothesis H₁ (Figure 4a). The Bayes Factor BF₀₁ = 0.246 indicates that the hypothesis of no difference in the number of stems between TRESTIMA and MOTI is less likely than the hypothesis of difference. The BF₀₁ error of 0.002% suggests that the BF₀₁ estimate is accurate. The median effect size
is 1.5, and the 95% credibility interval for the posterior distribution [0.289, 3.286], plotted as the line above the posterior distribution, indicates 95% certainty that the true population effect size is between 0.289 and 3.286. The fact that the circle on the posterior graph is much lower than the circle on the prior graph illustrates that the alternative hypothesis is favored. The Bayesian error probabilities of $P(H_0|\text{data}) = 0.20$ and $P(H_1|\text{data}) = 0.80$, calculated from Equation (1), indicate that the Bayesian error associated with a preference for $H_1$ is a remarkable 20%.

Figure 4. Bayesian paired sample $t$-tests for the difference between TRESTIMA and MOTI in (a) the number of stems and (b) stand basal area ($n = 4$ stands). The top panels show Bayes factors, medians of the effect size, and 95% credibility intervals. The middle panels show Bayes factor robustness. The bottom panels show how Bayes factors change with changing sample size and prior width.

The Bayes robustness check shows that the conclusion does not change at different prior widths (Figure 4a, middle panel). If we were less certain about the size of the effect and set a wider or an ultrawide prior width, the $BF_{01}$ would still indicate a moderate degree of belief in the alternative hypothesis, in the magnitude of 0.25 and 0.23, respectively. The sequential analysis (Figure 4a, bottom panel) confirms this stability.
The analysis shows mostly anecdotal evidence for H₁; for n = 4, the support in the data for H₁ is moderate, irrespective of the prior width. From the Bayesian analysis, we can conclude that our strength of belief that TRESTIMA and MOTI return different stem number estimates is moderate.

The frequentist paired sample t-test shows that the null hypothesis of no difference in stem number between TRESTIMA and MOTI should be rejected (t = 4.09, df = 3, p = 0.03; Table 3). Cohen’s d, which indicates the size of the difference between two means in standard deviation units, is large (Table 3). However, the Vovk–Sellke maximum p-ratio of 3.83 indicates that the p-value of 0.03 is only, at most, 3.83 times more likely to occur under H₁ than under H₀, which is clearly not strong evidence for H₁. Moreover, the frequentist Type I error probability α(p) (i.e., the frequentist probability in rejecting H₀) is 0.22, which means that for 22% of the paired sample t-tests, the p-value would be greater than 0.03 and thus more in favor of H₀. We could also say that in 78% of the t-tests, the results would be more significant than or equal to those in our case, and, in 22% of the tests, the results would be less significant.

### Table 3. Frequentist analysis of the differences in the number of stems and basal area between closest observations (n = 61) and the four study stands.

| Test         | Statistic | df  | p     | VS-mpr | Effect Size | 95% CI for Effect Size |
|--------------|-----------|-----|-------|--------|-------------|------------------------|
| TRESTIMA_N–MOTI_N | Student  | 4.09 | 3    | 0.03   | 3.83        | 2.04                   | 0.19  3.85          |
|              | Wilcoxon  | 1.83 | 0.13 | 1.42   | 1.00        | 1.00                   | 1.00  1.00          |
| TRESTIMA_G–MOTI_G | Student  | 2.28 | 3    | 0.11   | 1.54        | 1.14                   | –0.21 2.41         |
|              | Sign      | 1.5  | 0.13 | 1.42   | 1.00        | 1.00                   | 1.00  1.00          |
| TRESTIMA_G–MOTI_G | Student  | 2.52 | 60   | 0.01   | 6.00        | 0.32                   | 0.06  0.58          |

1. Vovk-Sellke maximum p-ratio: Based on the p-value, the maximum possible odds in favor of H₁ over H₀ equal 1/(−e p ln(p)) for p ≤ 0.37 [29]. For the Student t-test, effect size is given by Cohen’s d; for the Wilcoxon signed rank test and sign test, effect size is given by the matched rank biserial correlation. The sign test was used due to nonsymmetrical distribution of differences.

The nonparametric version of the paired sample t-test gives the opposite result. The significance of the exact nonparametric Wilcoxon signed-rank test (p = 0.13) suggests that we should retain the null hypothesis. When the asymptotic version of the test was used, the result was marginal (p = 0.068). However, these results should be interpreted with caution due to the low statistical power of tests with small samples. Therefore, the only conclusion that we can draw from the frequentist analysis is that neither p-values nor the two calibrated measures of the p-value (i.e., VS–MPR and α(p)) provide strong evidence for the rejection of H₀.

#### 3.2. Differences in Stand Basal Area between TRESTIMA and MOTI

Bayes Factor BF₁₀ = 0.729 for stand basal area indicates that the hypothesis of no difference between TRESTIMA and MOTI is only marginally less likely than the hypothesis of difference (Figure 4b, top panel). BF₁₀ = 0.729 is classified as anecdotal evidence in favor of the alternative hypothesis that the stand basal area differs between TRESTIMA and MOTI. The error of BF₁₀ was below 0.001%, which suggests that the BF₁₀ estimate is accurate. The median of the effect size was 0.866; the 95% credibility interval for the posterior distribution [−0.179, 2.178] indicates that the effect size is between −0.179 and 2.178. The Bayesian error probabilities of P(H₀|data) = 0.58 and P(H₁|data) = 0.42 suggest that if H₁ is preferred, there is still a 42% chance that H₀ is true.

The Bayes robustness check (Figure 4b, middle panel) shows that the degree of belief in either of the two hypotheses does not substantially change by changing prior width and that our weak belief in the alternative hypothesis should remain. The sequential analysis (Figure 4b, bottom panel) shows that a larger sample does not change our belief that neither of the two hypotheses is more likely.
The frequentist paired sample t-test shows that the hypothesis of no difference in stand basal area between TRESTIMA and MOTI should be retained ($t = 2.28$, df = 3, $p = 0.11$). Cohen’s $d$ is 1.141 (Table 3). The Vovk–Sellke maximum p–ratio of 1.54 indicates that the $p$-value of 0.11 is at most 1.54 times more likely to occur under $H_1$ than under $H_0$, which is clear evidence for $H_0$. The frequentist probability for rejecting $H_0$ $\alpha(p)$ is approximately 0.39, which means that for 39% of the paired sample t-tests, the result would be less significant and thus more in favor of $H_0$. The sign test, which we used as an alternative to the parametric test because of the small sample and skewness of the differences in $G$, also suggests retaining the null hypothesis. This result is somewhat counterintuitive since TRESTIMA returned consistently higher estimates of $N$ and $G$ for all stands.

The Bayesian paired sample t-test on the 61 nearest observations (i.e., without considering stand boundaries) shows results similar to its stand-wise counterpart. Bayes factor BF$_{10} = 2.533$ indicates that the odds of obtaining different stand basal area estimates with TRESTIMA are 2.533 to one (Figure 5, top panel). Different priors do not substantially change this belief (Figure 5, middle panel). However, since both TRESTIMA and MOTI provide point estimates, different standing point positions and camera directions may result in highly divergent estimates (Figure 5, bottom panel), which converge with a large enough sample. The frequentist test suggests that differences in $G$ between both applications, if subjective sampling is applied, are highly significant ($t = 2.52$, df = 60, $p = 0.01$; Table 3). This, however, indicates an uneven-aged stand structure rather than the poor accuracy of TRESTIMA. Using TRESTIMA, taking a few photos subjectively and hoping for an accurate population parameter estimate, may not be prudent unless large samples are used (Figure 5, bottom panel).

### 3.3. Informative Hypotheses

Evidence for the hypothesis of obtaining a greater number of stems versus a smaller number of stems with TRESTIMA compared to MOTI is extremely strong. The odds of overestimating versus underestimating $N$ for TRESTIMA compared to MOTI are 45,530 to 1 (Table 4). This means that there is a 99.9% probability of obtaining greater estimates of stand density than lower estimates with TRESTIMA compared to MOTI. The odds of obtaining a greater basal area with TRESTIMA compared to lower basal areas are 88.1 to 1. This means there is only a 1.1% probability of obtaining a lower basal area with TRESTIMA compared to MOTI. The analysis for a pooled sample ($n = 61$) shows roughly the same results.

| Hypothesis $^1$ | $n$ | BF   | Posterior Probability |
|----------------|-----|------|-----------------------|
| $H_2$: $N_{TRESTIMA} > N_{MOTI}$ | 4   | 45,529.8 | 0.9999                |
| $H_3$: $N_{TRESTIMA} < N_{MOTI}$ | 4   | 0.0000 |                       |
| $H_2$: $G_{TRESTIMA} > G_{MOTI}$ | 4   | 88.1  | 0.9888                |
| $H_3$: $G_{TRESTIMA} < G_{MOTI}$ | 4   | 0.0112 |                       |
| $H_2$: $G_{TRESTIMA} > G_{MOTI}$ | 61  | 168.5 | 0.9941                |
| $H_3$: $G_{TRESTIMA} < G_{MOTI}$ | 61  | 0.0059 |                       |

$^1$ Hypothesis $H_2$ specifies that the mean of TRESTIMA_N and TRESTIMA_G, respectively, is larger than the mean of MOTI_N and MOTI_G, respectively, while hypothesis $H_3$ specifies that it is smaller. The posterior probabilities are based on equal prior probabilities.
between both applications, if subjective sampling is applied, are highly significant ($t=2.52$, df = 60, $p=0.01$; Table 3). This, however, indicates an uneven-aged stand structure rather than the poor accuracy of TRESTIMA. Using TRESTIMA, taking a few photos subjectively and hoping for an accurate population parameter estimate, may not be prudent unless large samples are used (Figure 5, bottom panel).

**Figure 5.** Bayesian paired sample $t$-test for the difference in stand basal area between TRESTIMA and MOTI based on 61 pairs of closest observations.

### 4. Discussion

The performance of a mobile application can be evaluated from many perspectives. Software developers are concerned about accuracy, precision, and reliability, and the release of a new application is often accompanied by reported errors. However, inexperienced surveyors such as private forest owners may prefer to know whether a new application is as good as, better than, or worse than the
established one, be it in terms of accuracy, precision, reliability, or any other performance parameter such as price, simplicity, or speed. We showed that Bayesian reasoning can be more informative than the Neyman–Pearson approach when answering these questions. It has been said that people think in a Bayesian way [34]. Our real-life example supports this. Our private forest owner was not interested in whether the data supports the alternative hypothesis of a population difference between the two methods, but which one of the two hypotheses is more likely to be true after both applications “see” the data from his forest. In the frequentist framework, the answer to the latter question is not possible since the outcome of the statistical test cannot be that \( H_0 \) is accepted. High \( p \)-values do not mean that \( H_0 \) is true, and with a large enough sample size, \( H_0 \) will always be rejected [35]. Moreover, in the frequentist framework, one always assumes \( H_0 \) of no difference is true and subsequently seeks evidence against \( H_0 \). This means that any other plausible pair of hypotheses or their combinations cannot be compared. Bayesian hypothesis testing does not use thresholds such as significance levels but rather Bayesian factors, which measure the belief in the hypothesis, and are therefore much more intuitive for inexperienced end-users, who would probably have difficulty understanding the \( p \)-value.

On the whole, our analysis suggests that TRESTIMA estimates of the number of stems and stand basal area differ from MOTI estimates, but evidence for the differences between the applications is, at best, moderate. Conversely, there is almost no doubt that TRESTIMA overestimates stand parameters. The relative biases in basal area of +17.1% and +16.1% at the stand level and standing point level, respectively, is consistent with the results of Vastaranta et al. [10], who found a +15.2% relative bias. The relative bias for the number of stems of +49.4%, however, is inexplicably high. The reason for this could be the uneven-aged stand structure and relatively low number of trees for which tree diameter was recognized and used as an input in the stem number calculation [36]. The reason could also be that for Scots pine and other tree species, tree height was entered manually, while for spruce, it was calculated by the application. Overestimation of stand basal area leads to standing volume estimations that are too optimistic, which may have a negative impact on harvesting planning in small forest holdings. Moreover, imprecise basal area estimations due to small basal area factors may lead to bias in basal area increments [37] and subsequently to suboptimal harvesting levels. With respect to accuracy and precision, we would like to note that accuracy tests usually (and this is also the case in this paper) do not refer to the closeness of the measurements to the “true” population value of a stand parameter. Instead, estimates obtained by an alternative method are used as a reference (e.g., [11,38]), meaning that the measurement error of the reference method is not accounted for (but see [10,39,40]).

Our sample of four stands was extremely small for statistical hypothesis testing, meaning that the statistical power of the tests was low compared to the power usually used in the tests. Private forest owners should be warned about this before discussing the results. However, Bayesian tests are less influenced by sample size than frequentist tests, although small samples can also produce inconclusive results in Bayesian tests. In defense of the frequentist approach to statistical hypothesis testing with small samples, we would like to remind everyone that Student [41] used small samples of four units in his original paper on \( t \)-distribution. Moreover, many simulation and empirical studies suggest that the \( t \)-test can be applied to an extremely small sample size (\( n \leq 5 \)) as long as the effect size is expected to be large [42]. The core problem in real-life situations, however, is knowing the true effect size to estimate the Type II error. The effect size of an application can be estimated based on similar studies in the literature or expert opinion on the smallest effect size that is different enough from zero for the user [43]. However, users may have different criteria, and reported accuracy rates may be of little help for uneducated users. The positive side of the power analysis in the Bayesian framework is that the power analysis is not based on Type I and Type II error rates but on the probability that the Bayes factor exceeds the threshold value of the Bayes factor under the null hypothesis and under the alternative hypothesis [43]. In practice, power analysis means collecting new data and inspecting the Bayes factor until the sequential analysis diagram shows no substantial change. For the number of stems, the sequential analysis diagrams (Figure 4a, bottom panel) show that a large sample is needed for decisive evidence in favor of one of the hypotheses. Conversely, new data show similar weights for
the hypotheses about the basal area, i.e., the hypothesis of difference is only marginally more likely than the hypothesis of no difference (Figures 4b and 5, bottom panel), which means that the sample for achieving a sufficiently low probability of making an erroneous decision about the basal area is relatively small.

The Bayesian framework enables many competitive hypotheses to be evaluated, and analysts have flexibility in their formulation. For example, with the bain ANOVA [23], we can evaluate the plausibility of several claims about applications using equality and inequality constraints: $H_1$, that Application 1 provides greater estimates than Application 2, and Application 2 provides greater estimates than Application 3 ($H_1: \mu_1 > \mu_2 > \mu_3$); $H_2$, that all three applications are equal ($H_2: \mu_1 = \mu_2, \mu_3$); $H_3$, that Application 3 returns the highest estimate ($H_3: \mu_1 < \mu_2 < \mu_3$). However, if many hypotheses are formulated and evaluated, the Bayes factor will select the hypothesis that best describes the data and not, as it should, the hypothesis that best describes the population from which the data were sampled [22]. The conclusion may be more flawed when multiple hypotheses are tested on small samples; therefore, bain should not be misused for legitimizing uncontested beliefs about the applications.

Differences in stand basal area between both applications, when estimated on 61 pairs of closest observations, seem to vary substantially (Figure 5, bottom panel). However, the differences should be interpreted with caution. First, the standing point locations for TRESTIMA and MOTI were approximately 10 m apart because TRESTIMA samples trees in circular sectors while MOTI uses traditional proportional-to-size sampling in virtual circular plots. This means that the applications surveyed different parts of the forest. The differences may also originate from an uneven-aged stand structure, where a change in sampling point position may result in a greater change in the parameter than in even-aged forests. Vastaranta et al. [10] tested TRESTIMA in mainly even-aged and single layer stands, with an average size of slightly less than 1 ha, while in our case the stands were bigger and more variable in diameter. Second, the point-wise estimation of a stand parameter has no usefulness for a forest inventory. Bitterlich sampling selects trees with a probability proportional to their basal areas and thus provides local estimates of stand densities. A standing point level estimate is informative only when the variation of the parameter in the stand is close to null, such as in forest plantations. However, here, inventory crews should be cautious not to sample in similar locations relative to row spacing in order not to compromise the assumption of randomly selected point locations in the tract area [44]. Despite the differences found between both applications, our analysis does not indicate that TRESTIMA is inaccurate, but rather the opposite. We showed that a sufficiently accurate estimation of the basal area can also be obtained by nonprobability sampling, assuming that stand structure is not particularly uneven.

Owners of smaller forest holdings who use TRESTIMA in mixed deciduous forests may be less satisfied with the fact that TRESTIMA does not recognize all major tree species, does not estimate the number of stems per hectare for each individual standing point, and has relatively poorly documented computational processes [13]. For instance, the user has no influence on the determination of standing volume, but compared to MOTI, TRESTIMA enables the user to upload and store different types of GPS-tagged data and review the recorded photos with a computer or mobile device. It also enables stand characteristics to be shared with third parties as electronic reports or raw data.

5. Conclusions

We showed how private forest owners can use Bayesian statistics to evaluate a new mobile forest inventory application using a small sample and nonprobabilistic sampling. We believe that private forest owners or other inexperienced users are likely to have fewer problems understanding Bayesian hypothesis testing than the concept of null hypothesis significance testing, but only if the results of the Bayes tests are communicated in a simplified manner. We found differences between TRESTIMA and MOTI in the number of stems and stand basal area estimates; however, in the Bayesian sense, the differences were not large. Each private forest owner can decide how satisfied he is with the results of the application in his own forest and whether his testing contradicts our conclusions.
By collecting new data, the support for the hypotheses of interest can be continuously updated, and the reliability of our conclusions can be improved. However, private forest owners can already be sure that stand parameter overestimation with TRESTIMA is many orders of magnitude more likely than underestimation. The Bayesian approach could also be used in the multicriteria evaluation of the applications, where private forest owners decide the weights for certain criteria. Finally, no single or composite measure of application performance should replace commonsense reasoning.

Funding: This research was funded by the Ministry of Agriculture, Forestry, and Food of Republic of Slovenia, grant number 2330-17-000077, project FOREXCLIM, and the ARRS program P4-0059 “Forest, forestry and renewable forest resources”.

Acknowledgments: Metod Indihar performed the field measurements in the course of paid student work.

Conflicts of Interest: The author declares no conflict of interest. The funders had no role in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References
1. Molinier, M.; López-Sánchez, C.A.; Toivonen, T.; Korpela, I.; Corral-Rivas, J.J.; Tergujeff, R.; Häme, T. Relaspone-Mobile and Participative In Situ Forest Biomass Measurements Supporting Satellite Image Mapping. Remote Sens. 2016, 8, 869. [CrossRef]
2. Rosset, C.; Brand, R.; Caillard, I.; Fiedler, U.; Gollut, C.; Schmocker, A.; Weber, D.; Wuillemin, E.; Dumollard, G. MOTI. L’Inventaire Forestier Simplifié par le Smartphone. Version Actualisée du Rapport Final; Kessel Publishing House: Remagen, Germany, 2017; p. 12.
3. Mobile Forester. Available online: https://www.mobileforester.com/ (accessed on 25 February 2020).
4. KATAM™ Forest. Available online: https://www.katam.se/products/forest/ (accessed on 25 February 2020).
5. Liu, J.; Feng, Z.; Mannan, A.; Yang, L. Positioning of coordinates and precision analysis of sample trees using the intelligent forest survey calculator. Comput. Electron. Agr. 2019, 159, 157–164. [CrossRef]
6. Ficko, A.; Lidestav, G.; Ní Dhubháin, Á.; Karppinen, H.; Zivojinovic, I.; Westin, K. European private forest owner typologies: A review of methods and use. Forest Policy Econ. 2019, 99, 21–31. [CrossRef]
7. Bashir, A.; Sjølie, H.K.; Solberg, B. Determinants of Nonindustrial Private Forest Owners’ Willingness to Harvest Timber in Norway. Forests 2020, 11, 60. [CrossRef]
8. Trestima. Available online: https://www.trestima.com/en/ (accessed on 1 May 2019).
9. Rosset, C.; Brand, R.; Caillard, I.; Fiedler, U.; Gollut, C.; Schmocker, A.; Weber, D.; Wuillemin, E. MOTI–ein Tool für die Waldinventur im Taschenformat. Wald Holz 2015, 96, 45–48.
10. Rosset, C.; Brand, R.; Caillard, I.; Fiedler, U.; Gollut, C.; Schmocker, A.; Weber, D.; Wuillemin, E. MOTI-L’Inventaire Forestier Facilité par le Smartphone. Rapport Final. Projet no 2012.24; Haute école des sciences agronomiques, forêtières et alimentaires: Zollikofen, Switzerland, 2014; p. 110.
11. Turk, J. Analiza učinkovitosti aplikacij za pametne telefone pri merjenju sestojnih parametrov. Bachelor’s Thesis, University of Ljubljana, Slovenia, Biotechnical Faculty, Ljubljana, Slovenia, 2015.
18. Fricelj, B. Uporaba mobilne aplikacije MOTI za ocenjevanje sestojnih parametrov na zasebni gozdni posedi Lešje. Bachelor’s Thesis, University of Ljubljana, Biotechnical Faculty, Department of Forestry and Renewable Forest Resources, Ljubljana, Slovenia, 2018.

19. Rosset, C. Personal Communication; University of Applied Sciences: Bern, Switzerland, 2019.

20. Schütz, J.-P.; Rosset, C. Performances of different methods of estimating the diameter distribution based on simple stand structure variables in monospecific regular temperate European forests. *Ann. For. Sci.* 2020, 77, 47. [CrossRef]

21. Rouder, J.N.; Speckman, P.L.; Sun, D.; Morey, R.D.; Iverson, G. Bayesian t-tests for accepting and rejecting the null hypothesis. *Psychon. Bull. Rev.* 2009, 16, 225–237. [CrossRef]

22. Hoijtink, H.; Mulder, J.; van Lissa, C.; Gu, X. A tutorial on testing hypotheses using the Bayes factor. *Psychol. Meth.* 2018. [CrossRef] [PubMed]

23. Informative Hypotheses. Available online: https://informative-hypotheses.sites.uu.nl/software/bain/(accessed on 15 August 2020).

24. Zar, J.H. *Biostatistical Analysis*, 5th ed.; Pearson Prentice Hall: Upper Saddle River, NJ, USA, 2010; pp. 16–19.

25. Trestima. Trestima Forest Inventory System. User manual v1.4. Available online: https://trestima-www.s3-eu-west-1.amazonaws.com/docs/TRESTIMA_user_guide_en_v1.4.pdf (accessed on 25 August 2019).

26. Indihar, M. Primerjava mobilnih aplikacij MOTI in TRESTIMA® za oceno sestojnih parametrov. Bachelor’s Thesis, University of Ljubljana, Biotechnical Faculty, Ljubljana, Slovenia, 2020.

27. Wagenmakers, E.; Love, J.; Marsman, M.; Jamil, T.; Ly, A.; Verhagen, J.; Selker, R.; Gronau, Q.F.; Dropmann, D.; Boutin, B.; et al. Bayesian inference for psychology. Part II: Example applications with JASP. *Psychon. Bull. Rev.* 2018, 25, 58–76. [CrossRef] [PubMed]

28. Kass, R.E.; Raftery, A.E. Bayes Factors. *J. Am. Stat. Assoc.* 1995, 90, 791. [CrossRef]

29. Sellke, T.; Bayarri, M.J.; Berger, J.O. Calibration of p values for testing precise null hypotheses. *Am. Stat.* 2001, 55, 62–71. [CrossRef]

30. Wasserstein, R.L.; Lazar, N.A. The ASA Statement on p-Values: Context, Process, and Purpose. *Am. Stat.* 2016, 70, 129–133. [CrossRef]

31. Altman, N.; Krzywinski, M. Interpreting P values. *Nat. Methods* 2017, 14, 213–214. [CrossRef]

32. van Kesteren, E.J.; Wagenmakers, E.J. Exploring the diagnosticity of the p-value. Available online: https://www.shinyapps.org/apps/vs-mpr/ (accessed on 1 July 2020).

33. JASP Team. *JASP Version 0.11.1*; JASP Team: Amsterdam, The Netherlands, 2019.

34. Tégäs, E.; Vul, E.; Girotto, V.; Gonzalez, M.; Tenenbaum, J.B.; Bonatti, L.L. Pure reasoning in 12-month-old infants as probabilistic inference. *Science* 2001, 332, 1054–1059. [CrossRef]

35. Halsey, L.G. The reign of the p-value is over: What alternative analyses could we employ to fill the power vacuum? *Bio. Lett.* 2019, 15, 20190174. [CrossRef]

36. Kivimäki, S. Personal Communication; Trestima Oy: Tampere, Finland, 2020.

37. Eastaugh, C.; Hasenauer, H. Improved estimates of per-plot basal area from angle count inventories. *iForest Biogeo. For.* 2014, 7, 178–185. [CrossRef]

38. Siipilehto, J.; Lindeman, H.; Vastaranta, M.; Yu, X.; Uusitalo, J. Reliability of the predicted stand structure for clear-cut stands using optional methods: Airborne laser scanning-based methods, smartphone-based forest inventory application Trestima and pre-harvest measurement tool EMO. *Silva Fenn.* 2016, 50, 1568. [CrossRef]

39. Eastaugh, C.S.; Hasenauer, H. Biases in Volume Increment Estimates Derived from Successive Angle Count Sampling. *For. Sci.* 2013, 59, 1–14. [CrossRef]

40. Zhou, D.Q.; He, X.J.; Chen, G.W. Design and application of barcode diameter-at-breast-height tape in forest inventories. *Appl. Ecol. Envi. Res.* 2019, 17, 13407–13421. [CrossRef]

41. Student. The Probable Error of a Mean. *Biometrika* 1908, 6, 1–25. [CrossRef]

42. de Winter, J.C.F. Using the Student’s t-test with extremely small sample sizes. *Pract. Asses. Res. Eval.* 2013, 18, 10. [CrossRef]

43. Fu, Q.; Hoijtink, H.; Moerbeek, M. Sample-size determination for the Bayesian t test and Welch’s test using the approximate adjusted fractional Bayes factor. *Beh. Res. Methods* 2020. [CrossRef] [PubMed]
44. Lynch, T.B.; Hamlin, D.; Ducey, M.J.; Borders, B.E. Design-Unbiased Estimation and Alternatives for Sampling Trees on Plantation Rows. *Forests* 2018, 9, 362. [CrossRef]

**Publisher’s Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2020 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).