23rd International Congress on Modelling and Simulation, Canberra, ACT, Australia, 1 to 6 December 2019
mssanz.org.au/modsim2019

Uncertainty quantification in classification problems: A Bayesian approach for predicating the effects of further test sampling

Jordan Phillipson [a], **Gordon S. Blair** [a] **and Peter Henrys** [b]

[a] *School of Computing and Communications , Lancaster University, Lancaster,* [b] *Centre for Ecology and Hydrology, Lancaster*
Email:j.phillipson@lancaster.ac.uk

Abstract: The use of machine learning techniques in classification problems has been shown to be useful in many applications. In particular, they have become increasingly popular in land cover mapping applications in the last decade. These maps often play an important role in environmental science applications as they can act as inputs within wider modelling chains and in estimating how the overall prevalence of particular land cover types may be changing.

As with any model, land cover maps built using machine learning techniques are likely to contain misclassifications and hence create a degree of uncertainty in the results derived from them. In order for policy makers, stakeholder and other users to have trust in such results, such uncertainty must be accounted for in a quantifiable and reliable manner. This is true even for highly accurate classifiers. However, the black-box nature of many machine learning techniques makes common forms of uncertainty quantitation traditionally seen in process modelling almost impossible to apply in practice. Hence, one must often rely on independent test samples for uncertainty quantification when using machine learning techniques, as these do not rely on any assumptions for the how a classifier is built.

The issue with test samples though is that they can be expensive to obtain, even in situations where large data sets for building the classifier are relatively cheap. This is because tests samples are subject to much stricter criteria on how they are collected as they rely on formalised statistical inference methods to quantify uncertainty. In comparison, the goal of a classifier is to create a series of rules that is able to separate classes well. Hence, there is much more flexibility in how we may collect samples for the purpose of training classifiers. This means that in practice, one must collect test samples of sufficient size so that uncertainties can be reduced to satisfactory levels without relying overly large (and therefore expensive) sample sizes. However, the task of determining a sufficient sample sizes is made more complex as one also need account for stratified sampling, the sensitivity of results as unknown quantities vary and the stochastic variation of results that result from sampling.

In this paper, we demonstrate how a Bayesian approach to uncertainty quantification in these scenarios can handle such complexities when predicting the likely impacts that further sampling strategies will have on uncertainty. This in turn allows for a more sophisticated from of analysis when considering the trade-off between reducing uncertainty and the resources needed for larger test samples.

The methods described in this paper are demonstrated in the context of an urban mapping problem. Here we predict the effectiveness of distributing an additional test sample across different areas based on the results of an initial test sample. In this example, we explore the standard frequentist methods and the proposed Bayesian approach under this task. With the frequentist approach, our predictions rely on assuming fixed points for unknown parameters, which can lead to significantly different results and no formalised way to distinguish between them. In contrast, a Bayesian approach enables us to combine these different results with formalised probability theory. The major advantage of this from a practical perspective is that this allows users to predict the effect of an additional test sample with only a single distribution whilst still accounting for multiple sources of uncertainty. This is a fundamental first step when quantifying uncertainty for population level estimates and opens up promising future work in for the prorogation of uncertainty in more complex model chains and optimising the distribution of test samples.

Keywords: *Uncertainty quantification, land cover mapping, Bayesian, sampling strategies*

1. INTRODUCTION

Machine learning techniques are becoming increasingly popular tools in classification problems (Jordan and Mitchell 2015; Raut et al. 2017). The main advantages they bring over traditional process modelling is that they are able to make use of the complex interactions between many potential predictors and the target classes without the need for a high level of domain expertise (Kotsiantis 2007). The major drawback of machine learning techniques in classification problems is that they typically require many more labelled examples than traditional process models in order to be effective. In many applications, collecting enough samples randomly from an application space is simply not feasible. Hence, alternative methods are often needed to collect data sets to train classifiers (Pan and Yang 2010; Chawla et al. 2002). In such situations, uncertainty quantification based on the training data (e.g. statistical analysis on fitted parameter values, using the results from cross validation or out of the bag error, etc.) cannot be relied upon as the assumption that the data is representative of the application space is likely to be violated.

One way around this is to make use of test samples. Here we define a test sample as a set of labelled examples that is independent of any model building process and define test sampling as the act of collecting a test sample. Because of their independence from the model building stages, these test samples can then provide a universal approach to estimating and quantifying uncertainty for population level estimates (e.g. error rates, precision, recall, classes prevalence etc.) regardless of the type of classifier used or the data used to train them. However, test samples are subject to far stricter criteria than training sets and need to be collected in a much more statistically rigorous manner. Hence, for reasons similar to those that limit large random training samples, large test samples may also be impractical. This then often leads to a trade-off between cost and the degree of uncertainty in estimates when deciding upon the size of test samples.

From a high-level perspective, we wish to reduce the cost of test sampling by only collecting the necessary number of samples for a particular problem and decide how to make best use of the limited resources in test sampling in order to reduce uncertainty as much as possible. Specifically, this paper explores these trade-offs in the context of population level estimates from land cover mappings. Land cover maps have been shown to be useful in many environmental science applications including carbon emission monitoring (P. Olofsson et al. 2011; Avitabile et al. 2016), forest monitoring (Hansen et al. 2013; Townshend et al. 2012), modelling of soil properties (Shi et al. 2011) and biodiversity studies (Asner et al. 2009; Mendenhall et al. 2011).

A common approach in quantifying the uncertainty for population level estimates from a test sample is with the use of error matrices, a tabulation of the different types of errors seen in the test set. From this uncertainty for these estimates can then quantified with a frequentist perspective (i.e. with confidence intervals) (Pontus Olofsson et al. 2014). Practices such as stratified sampling can then be used to reduce uncertainty more efficiently than simple random sampling. Under the current frequentist approach, one can use a smaller initial test sample to estimate more efficient ways to distribute the remaining samples across the different strata and provide point estimates for the future measures of uncertainty (Wagner and Stehman 2015). The issue with a frequentist approach to uncertainty, however, is that it is difficult to account for multiple sources of uncertainty when estimating the effect of additional sampling.

In this paper, we investigate whether a Bayesian approach to uncertainty quantification for population level estimates can help overcome these challenges when it comes to analysing the effectiveness of different sampling strategies on the levels of uncertainty. Specifically, we evaluate these approaches in the context of an urban mapping problem.

2. NOTATION AND BACKGROUND

We begin with some definitions and notation. Suppose we have a total of c classes along with k strata. Next we define $\boldsymbol{p} := (\boldsymbol{p}'_1, \dots, \boldsymbol{p}'_k)'$ where $(\boldsymbol{p}_i)_j$ denotes the proportion of strata $i = 1, \dots, k$ that belongs to class $j = 1, \dots, c$. Next suppose we then collect a random samples from each strata of size $\boldsymbol{n} = (n_1, \dots, n_k)'$ (with replacement) from our strata and observe $\boldsymbol{x} = (\boldsymbol{x}'_1, \dots, \boldsymbol{x}'_k)'$ where n_i denotes the number of samples drawn from strata i and $(\boldsymbol{x}_i)_j$ denotes the number of observed instances from strata i $(i = 1, \dots, k)$ that belonged to class j $(j = 1, \dots c)$. In this situation, we can then model $\boldsymbol{x}$ as an observation drawn from a random variable $\boldsymbol{X}^* = (\boldsymbol{X}_1{}', \dots, \boldsymbol{X}_k{}')'$ where $\boldsymbol{X}_i$ follows a multinomial distribution with n_i trails and probability vector $\boldsymbol{p}_i$. We next define a population level quantity as any quantity that can be expressed in the form $g(\boldsymbol{p})$. Some popular population level estimates in machine learning applications are various performance metrics such as overall accuracy, sensitivity, specificity, precision etc. In addition, cost functions that take in to account different costs for different types of misclassifications are included in this generalised form. In land cover mapping applications, some other popular population level estimates are user and producer accuracies, area estimates and other quantities due to the land use change (e.g. total carbon emissions). The ultimate goal in this setting is to estimate $g(\boldsymbol{p})$ with the observed data $\boldsymbol{x}$. However, since $\boldsymbol{x}$ is only a sample of the entire population, there will inevitably be some uncertainty in such estimations. Typically, as the sample sizes become larger, the precision of our estimates will increase. Ideally though, one would want to make more sophisticated statements regarding this relationship as if one were able to achieve such a thing, one could then keep the cost of test sampling to a minimum by using sample sizes that are no larger than necessary.

However, the behaviour uncertainty is highly dependent on the unknown value of $\boldsymbol{p}$.To get around this, an ad-hoc approach is generally required when estimating the behaviour of uncertainty. Here one would collect test samples in batches rather than a single large sample with the possibility of remodelling our estimates of uncertainty and reconfiguring the sample distribution as one progresses through these batches. In this paper, we focus on the step between these batches. In terms of our original notation, we need to be able assess how a further set of test samples of size $\boldsymbol{n}^* = (n_1^*, \ldots, n_k^*)'$ is likely to effect the level of uncertainty for $g(\boldsymbol{p})$ given its current state of uncertainty with $\boldsymbol{x}$. More formally, suppose we have a way of measuring the degree of uncertainty for $g(\boldsymbol{p})$ given an observed sample,$\boldsymbol{x}$,which we denote with $\mathcal{A}(g(\boldsymbol{p}), \boldsymbol{x})|\boldsymbol{x}$ then we wish to be able to form an expression for

$$\mathcal{A}(g(\boldsymbol{p}), \{\boldsymbol{x}, \boldsymbol{X}^*\})|\boldsymbol{x} \qquad (1)$$

where $\boldsymbol{X}^* = (\boldsymbol{X}_1^*, \ldots, \boldsymbol{X}_k^*)$, $\boldsymbol{X}_i^* \sim \mathrm{Mult}(n_i^*, \boldsymbol{p}_i)$. Common examples of these uncertainty measures in a frequentist setting include standard error estimates and the length confidence intervals for a set level. In a Bayesian setting, any measure for the spread of a posterior distribution could act as a measure of the degree of uncertainty. Common examples of this would include the standard deviation of a posterior distribution or the length of a creditable interval at a set level.

3. PROBLEMS WITH A FREQUENTIST APPROACH

One way of quantifying uncertainty is to take a frequentist approach and use measures of uncertainty such as confidence regions (which are often simply confidence intervals in practice). Here the unknown value of $\boldsymbol{p}$ is assumed fixed and confidence regions are probabilistic statements made in relation to the test sampling, to which $\boldsymbol{x}$ is one instance of this sampling process. The use of confidence regions is currently the recommended practice within the land cover mapping community (Pontus Olofsson et al. 2014) and a standard approach in many other applications. By assuming a value for $\boldsymbol{p}$ one can then estimate how to best distribute an additional sample of fixed size across strata for simple (yet common) forms of g (Wagner and Stehman 2015). Different sampling distributions can then be compared under any fixed $\boldsymbol{p}$. The weakness of this approach is that it can only inform us of the behaviour at set points for $\boldsymbol{p}$. This makes it difficult to compare the likely impacts of different sampling distributions will have on the degree of uncertainty for estimates as they are dependent on the assumed value of $\boldsymbol{p}$. In terms of the notation, one can view this as only being able to give statements of the form

$$\mathcal{A}(g(\boldsymbol{p}), \{\boldsymbol{x}, \boldsymbol{X}^*\})|\{\boldsymbol{p}, \boldsymbol{x}\} \qquad (2)$$

This change in statement then leads to two major issues with a frequentist approach. The first issue is the problem of selecting an appropriate set of values for $\boldsymbol{p}$ to consider. One may generate confidence regions for components of $\boldsymbol{p}$ relevant to g from $\boldsymbol{x}$ as a guidance. However, this would be at best a heuristic approach to give a sensible range for values of $\boldsymbol{p}$ to consider. This is due to the fact that $\boldsymbol{p}$ is assumed to be fixed in a frequentist setting. Hence, distributions used to construct confidence regions are not formally a suitable measure to decide which values of $\boldsymbol{p}$ to prioritise. For example, suppose one uses a normal distribution centred $\widehat{\boldsymbol{p}}$ on to construct a confidence region for $\boldsymbol{p}$. Strictly speaking, it would be unsuitable to use the probability density function of this normal distribution as a measure to prioritise or assign weights to different values for $\boldsymbol{p}$.

The second issue that follows from this is in the difficulty of interpreting the results once an appropriate range for $\boldsymbol{p}$ has been agreed. For every fixed value of $\boldsymbol{p}$ used in (2) a distribution of values will be generated.

In practice this means one is often left with a large list of different distributions with no clear way of deciding which sampling distribution should be chosen.

4. AN ALTERNATIVE BAYESIAN APPROACH

An alternative approach to uncertainty quantification is with a Bayesian approach. The key difference here is that whilst $\boldsymbol{p}$ is still fixed and unknown, we allow the uncertainty for $\boldsymbol{p}$ given the observed data $\boldsymbol{x}$ to be represented as a probability distribution by using Bayes theorem. From this, one can then marginalise out $\boldsymbol{p}$ in (2) to give (1). The key consequence of this is that by being able to generate distributions in (1) rather than those in (2), this allows us to avoid the aforementioned issues under the frequentist approach. In terms of our notation, we begin with the probably density function for $\boldsymbol{p}|\boldsymbol{x}$, which is given by Bayes theorem

$$\pi(\boldsymbol{p}|\,\boldsymbol{x}) = \frac{\pi(\boldsymbol{x}|\boldsymbol{p})\pi(\boldsymbol{p})}{\pi(\boldsymbol{x})}$$

Where $\pi(\boldsymbol{x}|\boldsymbol{p})$ is the likelihood function, $\pi(\boldsymbol{x})$ is the marginal likelihood and $\pi(\boldsymbol{p})$ denotes a choice of prior distribution. In the context of stratified random sampling and proportion estimates, $\pi(\boldsymbol{x}|\boldsymbol{p})$ is derived by considering a series of likelihood functions for multinomial distributions. Since $\pi(\boldsymbol{x})$ effectively only acts as a normalising constant, the need to calculate $\pi(\boldsymbol{x})$ is avoidable in practice. The choice of a prior distribution $\pi(\boldsymbol{p})$ is perhaps the most controversial part of this procedure. Here we need to place a probability distribution to represent our belief in the likely values for $\boldsymbol{p}$ before observing $\boldsymbol{x}$. Determining suitable choices of priors is beyond the scope of this paper (although this issue is addressed in other fields of study (Berger et al. 2015)) . Rather, the purpose of this paper is to demonstrate a key advantage once a set of priors has been accepted. As a starting point, one may wish to consider the Dirichlet distribution (a multivariate extension to the beta distribution) for priors of $\boldsymbol{p}$ (Frigyik et al. 2010) as this distribution bring many advantages including the flexibility to fit a large range of priors, conjugacy, and the fact that this form of distribution includes many uninformative priors. For a more generalised prior structure, one may rely on well-known techniques such as Markov Chain Monte–Carlo (MCMC) sampling to generate posterior distributions (van Ravenzwaaij et al. 2018). However, it should be noted that these techniques can become computationally expensive in high dimensional settings. Regardless of the method used to generate a posterior distribution, once we have $\pi(\boldsymbol{p}|\,\boldsymbol{x})$ we can then express (1) with

$$\mathcal{A}(g(\boldsymbol{p}), \boldsymbol{x} + \boldsymbol{X}^*)|\boldsymbol{x} = \iint (\mathcal{A}(g(\boldsymbol{p}), \{\boldsymbol{x}, \boldsymbol{x}^*\})|\{\boldsymbol{x}, \boldsymbol{x}^*\})\, \pi(\boldsymbol{x}^*|\,\boldsymbol{p}), \pi(\boldsymbol{p}|\boldsymbol{x})\, d\boldsymbol{x}^* d\boldsymbol{p}$$

In practice, this is hard to solve analytically. Fortunately, one can generate a sample form $\mathcal{A}(g(\boldsymbol{p}), \{\boldsymbol{x}, \boldsymbol{X}^*\})|\boldsymbol{x}$ with a simple simulation precede: Step 1, generate $\boldsymbol{p}^*$ by drawing a sample of size 1 from $\boldsymbol{p}|\boldsymbol{x}$. Step 2, generate an artificial example $\boldsymbol{x}^*$ by drawing from a series of multinomial distributions with the assumption that $\boldsymbol{p} = \boldsymbol{p}^*$. Step 3, calculate $\mathcal{A}(g(\boldsymbol{p}), \{\boldsymbol{x}, \boldsymbol{x}^*\})|\{\boldsymbol{x}, \boldsymbol{x}^*\}$. This procedure can then be repeated multiple times to build a distribution for $\mathcal{A}(g(\boldsymbol{p}), \{\boldsymbol{x}, \boldsymbol{X}^*\})|\boldsymbol{x}$.

5. CASE STUDY: URBAN MAPPING IN LAGOS

Here we highlight the practical differences of a frequentist and Bayesian approaches when analysing the effect of future tests samples on uncertainty in the context of an urban mapping problem. We begin with a constructed land cover map for a region of Lagos and the surrounding area. In this scenario, the task is to estimate the total proportion of pixels that are represent urbanised areas as defined by a ground truth source. From this, one can then estimate the total urban area of the region. For convenience, this final step will be omitted in this paper. Table 1 provides a tabulation of an initial test sample of size 1000. Here pixels are randomly selected (with replacement) from each strata and are then manually assigned a class of either *Urban land*, *Nonurban land* or *Water*. The number of pixels drawn from each strata in this first sample is proportional to the area of the map each strata occupies. We then propose four new sampling distributions for the strata and analyse the likely effect each sampling strategy will have on the uncertainty for the true proportion of urban pixels should they be collected and combined with the original sample.

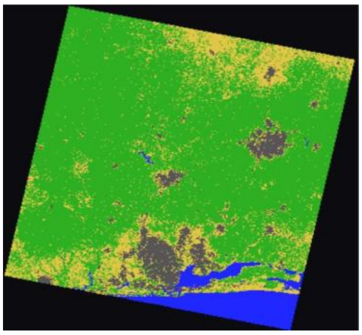

Figure 1. 2016 Urban mapping of the Lagos area. Key: Urban land (grey), Water (blue), Nonurban Land (yellow and green). Yellow Nonurban Lands indicate areas that are suspected as being more prone to errors.

Table 1. Error matrix obtained from the initial test sample along with the proposed distributions for each additional sample.

		Reference					Proposed Distribution for Next Sample			
	Predicted Class	**1**	**2**	**3**	**Total**	**Strata Size (W)**	**(i)**	**(ii)**	**(iii)**	**(iv)**
⬛	Urban Land (1)	56	0	7	63	0.063	63	344	900	189
🟦	Water (2)	0	60	1	61	0.062	61	0	50	183
🟨	Nonurban Land 1 (3)	10	0	163	173	0.173	173	655	50	519
🟩	Nonurban Land 2 (4)	1	0	702	703	0.703	703	0	0	2109

In terms of our notation, the total proportion of urban pixels can be expressed as $g_1(\boldsymbol{p}) := \sum_{i=1}^{4} W_i(\boldsymbol{p}_i)_1$ where $W = (0.063, 0.062, 0.173, 0.703)'$ and $\boldsymbol{x}$ represents the original test sample. Each new proposed sample is the equivalent of different $\boldsymbol{n}^*$ to analyse. The proposed distributions for each sample are as follows: (i) a further sample of size 1000 distributed in the same manner the original sample, (ii) a proposed optimised distribution for a further sample of size 1000 based on the assumption that $\boldsymbol{p} = \widehat{\boldsymbol{p}}$, (iii) a further sample of size 1000 with a heavy focus on sampling from the predicted urban class, (iv) a further sample of size 3000 distributed according to the strata size which was chosen post hoc to contextualise some results.

5.1 The use of frequentist methods

We consider the task of analysing and comparing the likely effects of the different sample distributions from (i)-(iv) when taking a frequentist perspective to uncertainty quantification. In order to be able to provide numerical examples, we shall use the estimated standard error of $g(\widehat{\boldsymbol{p}})$ as the choice of uncertainty measure in this example. The is we define $\mathcal{A}_F$ in this example as

$$\mathcal{A}_F(g_1(\boldsymbol{p}), \boldsymbol{x})|\boldsymbol{x} := \sqrt{\sum_{i=1}^{4} \frac{W_i^2(\widehat{\boldsymbol{p}}_i)_1(1-(\widehat{\boldsymbol{p}}_i)_1)}{n_i}}. \tag{3}$$

Where $\widehat{\boldsymbol{p}}$ denotes the maximum likelihood estimate (MLE) for $\boldsymbol{p}$ with $\widehat{\boldsymbol{p}} = \left(\frac{x_1'}{n_1}, \dots, \frac{x_k'}{n_k}\right)'$. The choice of measure in this case is not vital for the arguments presented in this paper but the estimated standard error is a reasonable choice as it is often used to create confidence intervals based on the asymptotic normality of $g_1(\widehat{\boldsymbol{p}})$. For example, a good approximation for a 95% confidence interval based on $\boldsymbol{x}$ in this case given by $g_1(\widehat{\boldsymbol{p}}) \pm 1.96 \times \mathcal{A}_F(g_1(\boldsymbol{p}), \boldsymbol{x})|\boldsymbol{x}$. In this case, the current standard error estimate from $\boldsymbol{x}$ is $\mathcal{A}_F(g_1(\boldsymbol{p}), \boldsymbol{x})|\boldsymbol{x} = 4.08 \times 10^{-3}$.

As discussed earlier, by taking a frequentist approach, we are only able consider the behaviour of future sampling strategies for fixed points of $\boldsymbol{p}$. Since only $(\boldsymbol{p}_i)_1,\ i = 1 \dots 4$ are needed in g_1 and $\mathcal{A}_F$ we only need to consider how these values vary in this case. For this case scenarios for the $(\boldsymbol{p}_i)_1 s$: (a) assumes each $(\boldsymbol{p}_i)_1$ is equal to its MLE, (b) and (c) are based on the tails of the individual 95% confidence intervals for each $(\boldsymbol{p}_i)_1$ and are designed to give pessimistic and optimistic estimates for the degree of uncertainty reduction achieved through (i)-(iv) respectively (see Figure 2 for more details). Here we are able the analyse the likely effects (i)-(iv) will have on the new value for $\mathcal{A}_F$ across each of assumed values in (a)-(c) through simulation based methods.

Proportion of Urban Pixels			
Strata	**MLE**	**Lower bound**	**Upper Bound**
Urban	0.8889[a]	0.7258[b]	0.9724[c]
Water	0.0000[a]	0.0000[c]	0.0578[b]
Nonurban 1	0.0578[a]	0.0281[c]	0.1037[b]
Nonurban 2	0.0014[a]	0.0000[c]	0.0079[b]

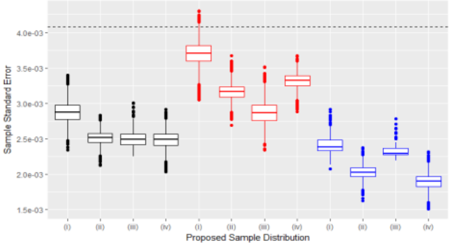

Figure 2. (Left) Summary table for the assumed rates in (a)-(c). Lower and upper bounds are derived from Clopper-Pearson intervals at the 95% level (Clopper and Pearson 1934). (Right) box plots of standard error estimates based on 1×10^5 simulations across the proposed sampling distributions and assumed rates [black = (a), red =(b), blue = (c)].

It is from here that we begin to see the weaknesses discussed in section 2 in practice. Across each assumed rate, results can vary significantly, both terms of the absolute degree of uncertainty reduction and the relative performances between (i)-(iv). In a frequentist setting like this, formally, there is no way to account for the relative plausibility of the values chosen in (a)-(c). This makes it difficult for users to decide which of these results provide better representations of the likely outcomes of the future sampling practices.

5.2 The use of a Bayesian approach

We next consider the task of analysing the likely effectiveness of future samples (i)-(iv) when taking a Bayesian approach to uncertainty quantification. With this different philosophical approach to uncertainty also comes a different way to measure the uncertainty. In this case we measure the degree of uncertainty given $\boldsymbol{x}$ as the standard deviation of the posterior distribution, $g(\boldsymbol{p})|\boldsymbol{x}$. Specifically, in this case we have

$$\mathcal{A}_B(g_1(\boldsymbol{p}),\boldsymbol{x})|\boldsymbol{x} := \sqrt{V(g_1(\boldsymbol{p})|\boldsymbol{x})} = \sqrt{\sum_{i=1}^{4} W_i^2 V((\boldsymbol{p}_i)_1|\boldsymbol{x}_i\ )}$$

For this case, we set a uniform prior on each $(\boldsymbol{p}_i)_1$ (i.e. we set $(\boldsymbol{p}_i)_1 \sim Beta(1,1)$). The choice of priors here is not vital to any fundamental arguments this paper makes. Merely, these priors were chosen based on a principle of indifference and to give similar numerical results to it frequentist counterpart, $\mathcal{A}_F$. From this we can then predict the likely effects each proposed future sampling distribution will have on $\mathcal{A}_B$ by following steps 1-3 in section 4 (although in this case, some of these simulation based steps are avoidable as closed forms for some distributions are easily obtainable). An example for this case is provided in Figure 3.

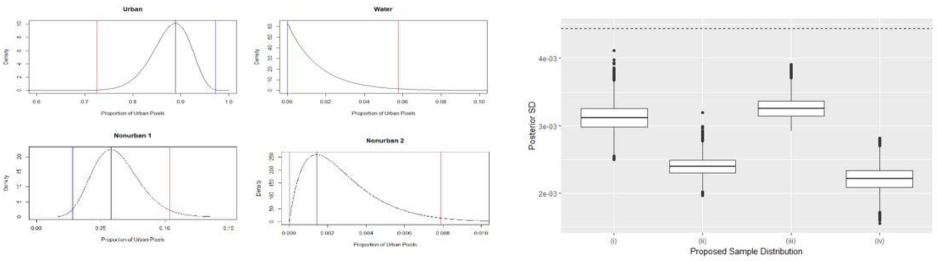

Figure 3. (Left) Posterior distributions for proportion of urban pixels within each strata from the initial test sample. Coloured lines indicate the rates assumed in Figure 2. (Right) box plots for the standard deviations of simulated posterior distributions based on 1×10^5 simulations across the proposed sampling distributions.

The key difference between the plots in Figure 2 and those in Figure 3 is that the credibility of any given scenario can calculated by using the posterior distribution for each of the $(\boldsymbol{p}_i)_1 s$ to assign weights. From this, we can calculate a weighted average across all scenarios to provide a single box-plot for each proposed sampling distribution. Furthermore, these weights are naturally taken in to account within simulation-based methods, lessening the burden of expertise required to apply this kind of analysis. From this, one is able to provide reliable predictions for the effects of future sampling distributions on uncertainty measures in a Bayesian setting, both in terms of relative and absolute effects. In this example specifically, we are able to see (i) and (iii) are likely to be less efficient than (ii) based on the information provided in the original test sample. Furthermore, there is likely to be a significant practical advantage in (ii) as the performance is similar to that of (iv) yet they are based on total sample sizes (i.e. including the original sample) of 2000 and 4000 respectively.

6 . CONCLUSION AND DISCUSSION

This paper has evaluated the suitability of different methods of uncertainty quantification when predicting the effects of further test sampling. Specifically, we then compared frequentist and Bayesian approaches in the context of an urban mapping problem. When predicting the effects of further test sampling, their impact on the degree of uncertainty is often governed by a set of unknown parameters, which themselves may need to be estimated with smaller initial test samples. When taking a frequentist approach to uncertainty quantitation, it is difficult to propagate uncertainty in these initial estimates. This can leave decision makers with a multitude of significantly different results with no formal framework to distinguish between them. In contrast, when using a Bayesian approach to uncertainty quantification on is able to propagate uncertainty from any initial estimates. This allows us to present the likely effects of a proposed test sample as a single distribution by formally combining the results with probability theory. This gives decision makers a method that is more reliable when: (1) assessing the trade-off between the cost of additional test sampling and the likely reduction in uncertainty for population level estimates, and (2) when assessing how different distributions of test samples across strata are likely to effect the efficiency of uncertainty reduction. Furthermore, many steps are achievable through simulation, making them simple to apply in practice.

This is promising but more experimentation is needed, including across other application domains and classification problems. It is also interesting to consider other advantages of Bayesian approaches, e.g. to support the propagation of uncertainty to other models or to support more adaptive sampling strategies when collecting testing data.

REFERENCES

Asner, G. P., S. R. Levick, T. Kennedy-Bowdoin, D. E. Knapp, R. Emerson, J. Jacobson, M. S. Colgan, and R. E. Martin. 2009. “Large-Scale Impacts of Herbivores on the Structural Diversity of African Savannas.” *Proceedings of the National Academy of Sciences*.

Avitabile, Valerio, Michael Schultz, Nadine Herold, Sytze de Bruin, Arun Kumar Pratihast, Cuong Pham Manh, Hien Vu Quang, and Martin Herold. 2016. “Carbon Emissions from Land Cover Change in Central Vietnam.” *Carbon Management*.

Berger, James O., Jose M. Bernardo, and Dongchu Sun. 2015. “Overall Objective Priors.” *Bayesian Analysis*.

Chawla, Nitesh V., Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. “SMOTE: Synthetic Minority over-Sampling Technique.” *Journal of Artificial Intelligence Research*.

CLOPPER, C. J., and E. S. PEARSON. 1934. “The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial.” *Biometrika* 26 (4): 404–13.

Frigyik, B., A. Kapila, and M. R. Gupta. 2010. “Introduction to the Dirichlet Distribution and Related Processes.” *Electrical Engineering*.

Hansen, M. C., P. V. Potapov, R. Moore, M. Hancher, S. A. Turubanova, A. Tyukavina, D. Thau, et al. 2013. “High-Resolution Global Maps of 21st-Century Forest Cover Change.” *Science*.

Jordan, M. I., and T. M. Mitchell. 2015. “Machine Learning: Trends, Perspectives, and Prospects.” *Science*..

Kotsiantis, S. B. 2007. “Supervised Machine Learning: A Review of Classification Techniques.” *Informatica (Ljubljana)*.

Mendenhall, C. D., C. H. Sekercioglu, F. O. Brenes, P. R. Ehrlich, and G. C. Daily. 2011. “Predictive Model for Sustaining Biodiversity in Tropical Countryside.” *Proceedings of the National Academy of Sciences*.

Olofsson, P., T. Kuemmerle, P. Griffiths, J. Knorn, A. Baccini, V. Gancz, V. Blujdea, R. A. Houghton, I. V. Abrudan, and C. E. Woodcock. 2011. “Carbon Implications of Forest Restitution in Post-Socialist Romania.” *Environmental Research Letters* 6 (4)..

Olofsson, Pontus, Giles M. Foody, Martin Herold, Stephen V. Stehman, Curtis E. Woodcock, and Michael A. Wulder. 2014. “Good Practices for Estimating Area and Assessing Accuracy of Land Change.” *Remote Sensing of Environment* 148: 42–57.

Pan, Sinno Jialin, and Qiang Yang. 2010. “A Survey on Transfer Learning.” *IEEE Transactions on Knowledge and Data Engineering*.

Raut, Priyanka P, Namrata R Borkar, Me Student, Assistant Professsor, and Sau Kamlatai. 2017. “Machine Learning Algorithms:Trends, Perspectives and Prospects.” *International Journal of Engineering Science and Computing*.

Ravenzwaaij, Don van, Pete Cassey, and Scott D. Brown. 2018. “A Simple Introduction to Markov Chain Monte–Carlo Sampling.” *Psychonomic Bulletin and Review*..

Shi, Wenjiao, Jiyuan Liu, Zhengping Du, Alfred Stein, and Tianxiang Yue. 2011. “Surface Modelling of Soil Properties Based on Land Use Information.” *Geoderma*..

Townshend, John R., Jeffrey G. Masek, Chengquan Huang, Eric F. Vermote, Feng Gao, Saurabh Channan, Joseph O. Sexton, et al. 2012. “Global Characterization and Monitoring of Forest Cover Using Landsat Data: Opportunities and Challenges.” *International Journal of Digital Earth*..

Wagner, John E., and Stephen V. Stehman. 2015. “Optimizing Sample Size Allocation to Strata for Estimating Area and Map Accuracy.” *Remote Sensing of Environment*.