Managing uncertainty in machine learning techniques: An investigation of adaptive sampling strategies through land cover mappings

Jordan Phillipson

School of Computing and Communications Lancaster University

This dissertation is submitted for the degree of
Doctor of Philosophy

March 2024

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this thesis are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This thesis is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the contribution statements at the beginning of this thesis. This thesis does not exceed the maximum permitted word length: it contains fewer than 80,000 words including appendices and footnotes but excluding the bibliography.

Jordan Phillipson

March 2024

List of Publications

The following has been published as part of the research presented in this thesis. Where appropriate, portions of this thesis are based on my contributions to these publications without citation. Where research and text should be credited to a co-author, rather than myself, the work has been cited accordingly.

Phillipson, J., Blair, G., & Henrys, P. (2022). Quantifying uncertainty in land cover mappings: An adaptive approach to sampling reference data using Bayesian inference. Environmental Data Science, 1, e15. https://doi.org/10.1017/EDS.2022.14

Phillipson, Jordan, Gordon Blair, and Peter Henrys. 2020. "Quantifying Uncertainty for Estimates Derived from Error Matrices in Land Cover Mapping Applications: The Case for a Bayesian Approach." In *IFIP Advances in Information and Communication Technology*.

Phillipson, J., G.S. Blair, and P. Henrys. 2019. "Uncertainty Quantification in Classification Problems: A Bayesian Approach for Predicting the Effects of Further Test Sampling." In *MODSIM2019, 23rd International Congress on Modelling and Simulation*, ed. S Elsawah. Canberra: Modelling and Simulation Society of Australia and New Zealand.

Acknowledgements

Firstly, I would like to thank my supervisors, Gordon Blair and Peter Henrys. You two have been unbelievably supportive (and patient) throughout. I cannot thank you both enough. In addition, I would like to thank my family, friends, and the wider Ensemble team for their support during this thesis.

Abstract

In recent decades, the use of machine learning techniques in classification problems has become increasingly popular across a wide variety of domains. For users to have trust in such classifiers though, one must be able to reliably quantify uncertainty. A common way of quantifying uncertainty in classifiers is through reference sampling where a smaller set of ground-truths is sampled and compared to their predicted counterparts to make inferences about the precision and accuracy of classifiers using statistical methods.

However, classification via machine learning can bring some additional challenges to uncertainty quantification, as machine learning techniques are often (i) trained using data that has not been sampled with formal statistical inference in mind; (ii) are often black-box when compared to traditional modelling.

These issues are further compounded when sampling reference data under conditions suitable for uncertainty quantification is expensive. Here, users are often forced to make a compromise between the degree of uncertainty and the costs of reference sampling, even when the original classifier built using machine learning may be performing well. In short, when it comes to quantifying and reducing uncertainty, it is not just about how well the classifier performs. One must also be able to collect enough data sampled under the right conditions.

This thesis explores how users may better manage the cost-benefit trade-offs of reference sampling when quantifying and reducing uncertainty in machine learning classifiers. Specifically, this thesis investigates how a framework for adaptively sampling reference data can be used to better manage uncertainty using two land cover mapping case studies to evaluate the proposed framework. With these case studies, the following problems are considered: (i) quantifying uncertainty in area estimation and mappings; (ii) proposing efficient sample designs under uncertainty; (iii) proposing sample designs when the cost of reference sampling varies across a mapped region.

Table of Contents

Chapter 1: Introduction

1.1 Motivation

The act of taking a population and separating its members into meaningful categories is a long-established practice in scientific applications. This can range from low stake applications such as separating spam emails from more genuine ones [1], [2], [3] or determining the breed of a dog from an image [4], [5], [6], to much higher stake applications such as categorising which people are suffering from a particular disease or illness [7], [8], [9], [10] or determining if a credit card transaction is likely to be fraudulent [11], [12], [13].

In many applications though, it is not practically possible to place populations into categories manually. This can be because of the size of the population, or the costs involved in categorising each member of the population. For example, there may be too many emails being sent at any given moment for a team of people to decide what should be considered spam email. Similarly, fraudulent credit card transactions are a relatively small percentage of the total transaction amounts [14], so a thorough review of every transaction would likely cost more than they would lose to fraud in the first place.

In cases when manually categorising large portions of a population is not viable, it is often beneficial to make use of classification algorithms, which this thesis will refer to as *classifiers* as a shorthand. In this context, classifiers are algorithms that aim to automate categorisation. Typically, these instructions involve a series of inputs, which this thesis will refer to as predictors for reasons that should become clearer later. Classifiers can be either discrete or fuzzy in their application. A *discrete classifier* is one where the classifier aims to place members in exactly one of the available categories. A *fuzzy classifier* is one where members may partially belong to multiple categories (for example a dog may be a mix of different breeds).

One approach to constructing classifiers that has become popular in recent times is to use machine learning techniques (MLTs). The distinction between MLTs and other forms of modelling lies in the more automated nature of MLTs, which often makes them a much more scalable approach when dealing with large data sets.

Classifiers (even those built with MLTs) will rarely be perfect though, and one will always need to account for such imperfections. One way of accounting for imperfections in classifiers is through *uncertainty quantification* (UQ). In this thesis, UQ refers to a process of providing a probabilistic statement as a measure of how confident one is in the true value of an unknown quantity based on a prediction provided by a classifier. UQ can capture this confidence at different levels, with some measures reflecting an aggregate level of confidence (e.g., overall accuracy) and other measures reflecting confidence for the classification of a single member of the population.

In most cases, methods of uncertainty quantification require some form of a *reference sample*. Here, a reference sample refers to a set of data that is collected in a specific way (i.e. under a sample design) that allows one to estimate unknown parameters used to express uncertainty in quantifiable terms. For example, one will typically not know how accurate a classifier will be when applied to a target population, but it is possible to estimate this overall accuracy value using a randomly selected subset of the population (i.e. simple random sampling). The uncertainty in this estimate may be quantified by taking advantage of the simple random design and well-known results in statistical inference (e.g. using confidence intervals).

In a perfect world, one would always have large sets of data collected under simple random sampling when quantifying uncertainty, as simple random sampling has many properties that make UQ through statistical inference much easier to deal with (Section 2.2 will go into more depth on this topic). However, practical restrictions often make simple-random sampling at a large-scale unviable. This problem is indicative of the wider problem of sampling design where meeting the design requirements necessary for formal statistical inference can make reference sampling expensive (even when large sets of training data are available).

As an example, suppose one is wanting to quantify uncertainty under a machine-learning classifier in a fraud detection application where manual categorisations may require lengthy and expensive investigations. Here, one may be able to build an accurate classifier using a set of training data containing confirmed fraudulent cases that have accumulated naturally over many years. Within this data set though, there is likely to be some bias in which members have been manually categorised. For instance, it may not be cost-effective to conduct investigations on low-value cases, and clients may not appreciate being inconvenienced (or worse, openly suspected of fraud) with an investigation given little initial evidence. Hence, this original reference data based on past cases is unlikely to be representative of the entire population. Unless one can explicitly quantify how this process is defined (which is likely to be difficult when such decisions involve human judgement), many methods of UQ that are based on statistical inference will be inapplicable when attempting to use this original training set.

This can create unfortunate situations where MLTs may offer reliable and cost-effective ways of creating accurate classifiers but are still left imprecise because one does not have enough of the right types of data to justify more precise estimates. In other words, predictions from the MLTs are forced to remain imprecise not because the MLTs are necessarily doing a poor job at classification, but because one does not have enough data sampled under the right conditions to formally justify higher levels of precision.

The overall motivation of this thesis is to address these kinds of situations by better managing the trade-off between sampling costs and uncertainty when dealing with classifiers built using MLTs. Whilst there is much in the literature for managing uncertainty in traditional modelling settings (e.g. power analysis, targeted sampling etc.), the use of MLTs brings a number of additional challenges, some of which include:

(i) Many MLTs are not designed to include formal uncertainty quantification.

(ii) MLTs often need to be trained on reference data that is not suitable for UQ as this may be the only way of generating large enough training sets.

(iii) UQ often involves many subjective choices related to modelling and sampling assumptions. These assumptions can be particularly hard to verify in MLTs as they often lack interpretability and explainability.

(iv) The field of machine learning, in general, is one of regular change that draws from a broad range of philosophies and ideas, meaning that MLTs considered state-of-the-art today may look very different from the state-of-the-art MLTs several years later.

The major consequence of these challenges is that much of the current literature for managing uncertainty efficiently does not easily translate over to MLTs. Hence, a key part of addressing the overall motivation lies in developing a set of methods that can deal with such complications. In particular, this thesis considers how such challenges may be overcome through a framework that (i) samples the data used in UQ adaptively, and (ii) focuses on methods that are agnostic to the choice of MLT, how it has been trained, or the specific method of UQ.

1.2. Background

1.2.1. Definitions and terminology

This thesis focuses on how uncertainty can be managed efficiently in classification problems involving machine learning techniques. Ultimately, this work lies at the intersection of three topics: machine learning techniques, uncertainty quantification and reference sampling (as illustrated in Figure 1.1). This subsection sets out what is meant by these three terms before setting out what is meant by managing uncertainty efficiently.

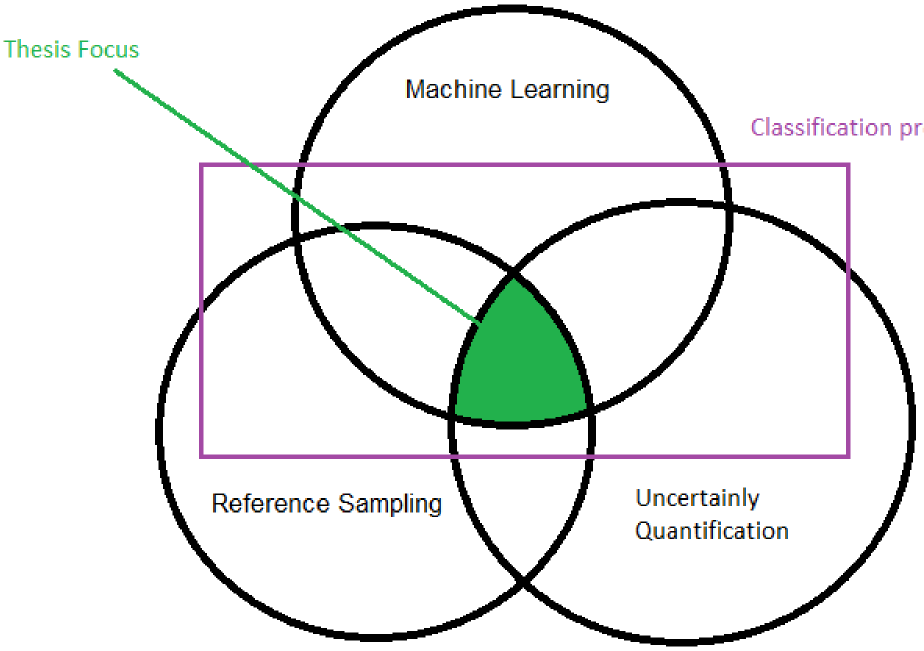


Figure 1.1. *A pictorial representation for the focus of this thesis.*

Machine learning techniques

A machine learning technique (MLT) for this thesis is an algorithm that attempts to automate the process of learning. The vagueness of this definition (in particular terms such as “automated” and “learning”) reflects how widespread machine learning has become since it was first coined in the 1950s [15]. Early applications of machine learning in the 1960s often focused on pattern recognition to learn effective strategies in well-defined gaming situations [16], [17], [18], [19]. Over the past 70 years though, machine learning has expanded to many other domains including medical diagnosis [20], [21], [22], image recognition [23], stock trading [24], [25], [26], [27], and climate modelling [28], [29], [30]. With this, machine learning has arguably evolved to become a crucial sub-discipline of data science. This evolution can be summarised with machine learning developing from automatically answering questions such as “What is the best move to play in this game?” to also including questions such as “What is the relationship between these collections of variables?”.

Because machine learning has entered so many domains over the years, it is difficult to give a precise definition of a machine learning technique that fully captures these different applications without including almost all forms of modelling. For example, one commonly cited definition for a machine learning technique given by Mitchell [31] is:

"A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E."

With this definition though, any model in which the parameters are estimated using sampled data could be considered a machine-learning technique. This would go very much against the spirit of machine learning, which is built on the idea of a computer programme extracting the relationships automatically, rather than starting with the relationship and using data to refine the specifics of such a relationship.

For example, suppose two variables are linearly related. An MLT could discover this relationship automatically (once given enough data), whilst a more traditional modelling approach may assume this linear structure initially and use data to estimate the most appropriate choice of parameters (in this case, the gradient and intercept). Given enough data, both approaches may reach a similar conclusion, but machine learning starts from a different level of assumed knowledge.

When exactly a method of building a classifier becomes an MLT can be up for debate. For example, it is not clear when semi or nonparametric modelling crosses over into machine learning. Conversely, within any MLT, there is always a degree of domain knowledge necessary when some of the parameters or features need to be set by the users (e.g. loss functions, smoothing parameters, defining how one will judge two instances as similar etc.). Hence, deciding when a method is so sensitive to these choices that it can no longer be considered an MLT can be a subjective matter.

Consequently, a precise definition of machine learning is set aside for this thesis. Instead, the term MLT is used as a descriptor that summarises a type of method which tends to have the following characteristics:

- In general, MLTs place less reliance on knowing the physical processes involved in a system when compared to traditional forms of modelling.
- MLTs often need large volumes of data to be effective.
- Many popular MLTs lack the interpretability and explainability seen in more traditional modelling. In other words, there is a tendency for MLTs to be black-box in nature.

It is important to stress here that these characteristics are not a definitive set of properties that must be observed for a method (classifier or otherwise) to be considered an MLT. For example, some classification methods are commonly called MLTs and are interpretable or effective with low volumes of data (these methods will be reviewed in detail throughout Chapter 2). Nevertheless, these characteristics are useful generalisations when trying to discuss the unique challenges MLTs bring to managing reference sampling and uncertainty quantification.

Uncertainty quantification

Uncertainty quantification (UQ) in this thesis refers to the act of providing a probabilistic statement to express confidence in estimations or predictions based on a current level of information. Again, this definition is left deliberately vague as a means of recognising that there are many viable approaches to quantifying uncertainty (see Section 2.1 for further details). At a high level, these different approaches can rely on subtly different assumptions and perspectives about how one interprets a set of data. Some examples of this include whether to treat unknown values as fixed or to allow them to partially belong to multiple values; assumptions about the specific way in which any data have been obtained; and whether one believes particular modelling assumptions are appropriate. This, in part, motivates the term uncertainty *quantification* as opposed to more deterministic language such as uncertainty *calculation*.

To compare different methods of UQ, it can be useful to consider uncertainty as a combination of multiple components of uncertainty. For this thesis, uncertainty will be expressed as a combination of three components, aleatoric, epistemic, and ontological uncertainty [32]:

- *Aleatoric uncertainty* - derived from the Latin *aleator* or *dice player* and is sometimes referred to as irreducible uncertainty. It is this form of uncertainty that cannot be reduced through sampling. In practice, aleatoric uncertainty is commonly expressed as a purely stochastic process. Here, it is often used to capture things such as measurement errors or to represent a “remainder” component in a model.

- *Epistemic uncertainty* - based on the Greek word for knowledge, episteme, epistemic uncertainties are components of the uncertainty that can in principle be known within a fixed system. It is this form of uncertainty that we expect to quantifiably reduce through sampling. Sources of epistemic uncertainty include uncertainty in the true value of model parameters or uncertainty in the true values of input features that have themselves been estimated as part of a wider modelling chain.

- *Ontological uncertainty* - ontological uncertainty arises from different beliefs regarding the true nature of a process. This could be things such as the belief of whether particular modelling assumptions are realistic or if the sample size is sufficient for one to take advantage of well-known results such as the central limit theorem. Evidence supporting underlying assumptions can help reduce ontological uncertainties, which may make use of reference data. This could be with simple visual validation or with more sophisticated statistical testing.

To illustrate components of uncertainty, suppose one has a simple linear regression model between two observations, x_i and y_i,

$$y_i = \mathrm{\beta}_0 + \mathrm{\beta}_1 x_i + \mathrm{\epsilon_i}, \mathrm{\epsilon_i} \sim \mathrm{N}(0, \mathrm{\sigma}^2). \tag{1.1}$$

When quantifying the uncertainty in some y_i given x_i in this example, the three sources of uncertainty can be observed. Firstly, the noise component, $\mathrm{\epsilon_i}$, generates aleatoric uncertainty. Secondly, uncertainty in the values of $\mathrm{\beta}_0, \mathrm{\beta}_1, \sigma$ are sources of epistemic uncertainty. Finally, the question of whether the relationship between y_i and x_i is linear as well as the structure of $\mathrm{\epsilon_i}$ (e.g. questions over whether they are Gaussian and independent) are sources of ontological uncertainty.

It is also important at this stage to differentiate between the concepts of *uncertainty* and *ambiguity* and how they relate to UQ and *fuzzy classification*. In short, uncertainty in classification problems refers to the degree to which a particular value is unknown, which is quantified using probabilistic statements. Ambiguity refers to situations when instances do not belong to exactly one pre-defined class and so need to be manually categorised using fuzzy logic rules. In the same way that a discrete classifier aims to emulate manual categorisation, fuzzy classifiers aim to emulate or estimate these fuzzy logic rules. Both uncertainty and ambiguity may be present within a given problem. Example 1.2 illustrates the difference between uncertainty and ambiguity.

Example 1.2*. An illustrative example of the difference between uncertainty and ambiguity under classifying dog breeds from images.*

Example: Uncertainty and Ambiguity

Suppose one wishes to classify the breed of a dog from the images below. For the sake of simplicity, assume that the dog is some mixture of an American bulldog and Staffordshire bull terrier (with the possibility of being a pure breed of either case)

Here, the difference between uncertainty and ambiguity can be exemplified by the statements in the table below.

	No ambiguity	Ambiguity is present
No uncertainty	The dog is known to be a purebred American bulldog.	The mother is known to be a purebred American bulldog and the father is known to be a purebred Staffordshire bull terrier (i.e., known to be 50-50).
Uncertainty is present	The dog is known to be either a pure-breed American bulldog or a pure-breed Staffordshire bull terrier, but one can not tell for sure which one.	The dog is a mixture of American bulldog and Staffordshire bull terrier (including purebred) but the exact degree is not known.

Reference sampling.

In this thesis, a reference sample consists of two components, the reference data and the sample design. Reference data are a record of information that one believes may be relevant to a specific classification problem. For example, the reference data may be a set of images that have been manually categorised, or it may be a set of results from lab testing, where each result is paired with a set of potential predictors. The sample design is a description, or set of instructions, that dictates how the reference data are (or were) obtained. Often, there will be stochastic elements used in the sampling design. Strictly speaking though, a stochastic component is not necessary, but it is often required for UQ or to avoid problems such as overfitting or bias when training MLTs. One of the most well-known sample designs is simple random sampling.

It is important to emphasise that both the reference data and sample design are necessary if one wishes to use a reference sample for most methods of UQ. Firstly, the reference data are needed to provide the raw numerical inputs. The reason the design is also needed is that many methods of UQ are based on precise probabilistic statements, which relate the reference data, sample design, and true values of unknown parameters in some way. Typically, these statements follow a structure similar to "given that one has observed this reference data under this sample design, it is likely that the true value for this unknown parameter is within this range". Note, this means that the uncertainty quantified from two identical sets of reference data may be quite different if they have been obtained from different sample designs.

As an example, one can consider a scenario where an amusement park wishes to know how satisfied users are with the park (e.g. opinion of the overall cleanliness, if attractions are easy to find, queuing times etc.). In this case, the manual categorisations are the responses given by the visitors to related questions in a questionnaire. Naturally, interviewing all the guests is not a realistic prospect, so one may need to estimate the true distribution of the responses through a survey. The responses to such a survey make up the reference data. If the participants are selected randomly, one may make use of straightforward statistical approaches to quantify uncertainty in such estimates. Suppose though, that instead of a random sample, the interviewers only ask the guests in the queue for the main attraction during peak hours for the sake of convenience (e.g. people may be more willing to answer a small survey whilst queueing to pass time). With the reference data from this design, quantifying uncertainty becomes more difficult than in the previous case. This is because there are biases in the design that one needs to account for when quantifying uncertainty. For example, those already in a long queue during peak hours are likely to be more tolerant of queuing times by the fact that they willingly entered such a queue. Other biases may be less direct. For example, the main attractions in an amusement park are often one of the fastest or largest in the park and typically have a higher minimum height requirement. This can be a proxy source of bias for other factors such as age or gender.

In general, one will often encounter the problem where some sample designs offer a more convenient way for collecting the reference data, but at the expense of requiring additional steps or assumptions to be made when quantifying uncertainty (and vice-versa). Choosing the most appropriate sample design in these situations often means balancing the trade-offs between convenience, cost, and the assumptions necessary for UQ.

Managing uncertainty efficiently.

For this thesis, managing uncertainty efficiently refers to quantifying and reducing uncertainty in a way that makes the best use of limited resources. Some examples of efficient uncertainty management include:

- Optimising designs in reference sampling to give more precise estimates for a set level of resources.
- Using a cost-benefit analysis to decide an appropriate level of resources to reach a desired level of uncertainty.
- Sampling reference data in a way that allows for a similar level of precision in estimates whilst reducing reliance on assumptions in UQ (i.e. a reduction in ontological uncertainty without a noticeable loss in precision).
- Exploiting new ways of generating reference data to make reference sampling cheaper.
- Using state-of-the-art classification methods to improve the quality of predictions and reduce uncertainty.

The important thing to note here is that for this thesis, efficiency in uncertainty management does not solely focus on optimising reference sampling under some predefined objective functions. Instead, efficiency in the context of this thesis relates to using the best combination of tools from machine learning, UQ, and reference sampling so that uncertainty is quantified and reduced effectively.

1.2.2 Domain of application: land cover maps

As part of the methodology this thesis will make use of case studies within land cover mapping applications to develop and evaluate methods related to adaptive sampling (see Section 1.3 for further details). Within land cover mapping applications, there are two key components to consider, the land cover map itself, and ground-truth assessments. A land cover map is a spatial representation of how the surface of a landmass varies. This can include things such as different types of vegetation cover, the degree of urbanisation, inland water, bare soil etc.

A ground-truth assessment, by definition, is the most accurate (or precise) categorisation of land cover available for an area. Some examples of ground-truth assessments include physical surveys, assessments based on lab testing or local sensory data, and assessments based on high-resolution aerial imagery (e.g., using aerial photography obtained from a drone). These assessments may be discrete (e.g., assigning an area based on the dominant type of land use) or continuous (e.g., a proportional breakdown of the different types of land cover across a large area).

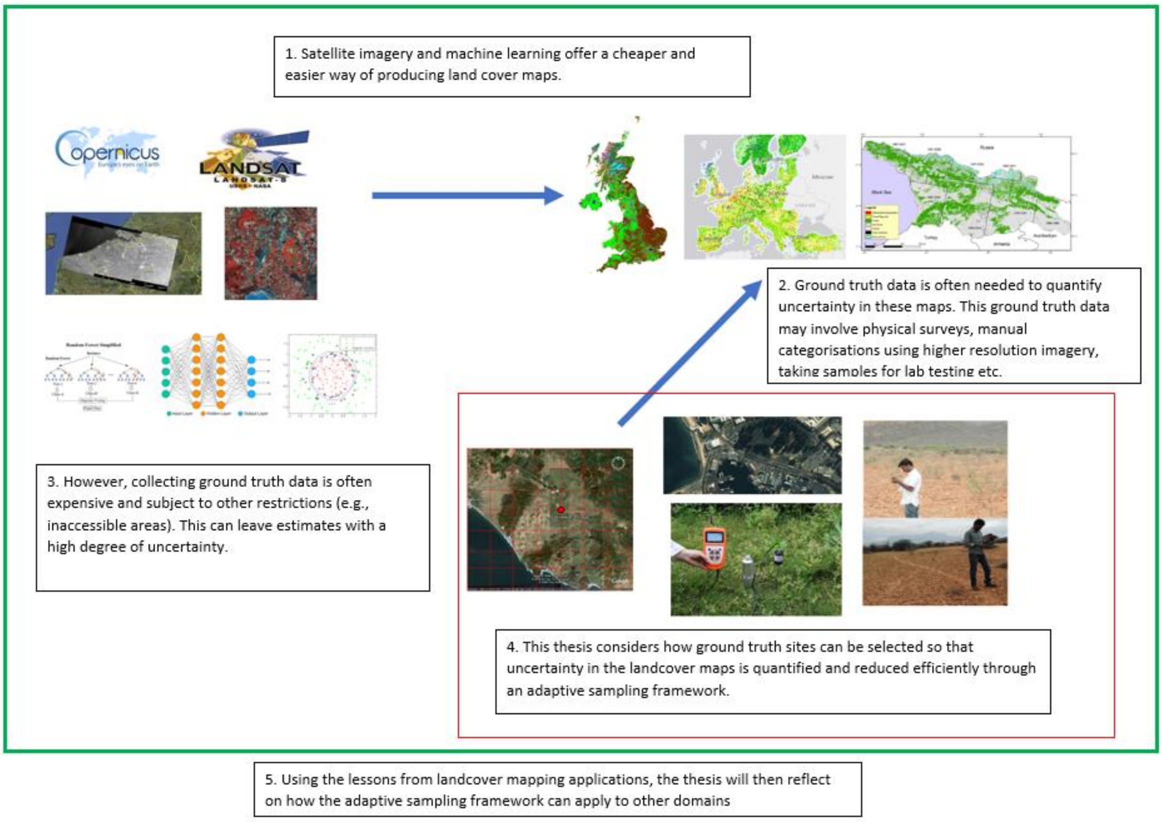
As a side note, ground-truth assessments are not necessarily objective assessments. For example, two surveys may come to slightly different assessments of the same area and there is usually some form of noise due to measurement error in any sensing equipment. Theoretically, non-objective ground-truths could be accounted for in modelling through additional aleatoric components. However, this thesis will assume all ground-truths to be objectively true as i) non-objective ground-truths go beyond the scope of the thesis; ii) to explore this concept fully, one would need to have access to case studies where multiple ground-truth assessments of the same areas are available, and such examples are not commonly found in land cover mapping applications and it would be expensive to generate such data.

Ideally, land cover maps would be built using full coverage census – i.e. made purely from ground-truth assessments. However, this tends to be unrealistic as applications often cover national or multinational areas. When this is the case, land cover maps serve to model an area and will act as the classifiers in the context of this thesis.

Since the turn of the 21st century, there has been a growing trend of using satellite imagery data and MLTs to construct land cover maps [33], [34], [35]. The main idea here is that satellite imagery acts as an inexpensive source of data that covers the entirety of a mapped area, making it a popular source of predictive features in classifiers. MLTs then provide a way of using the features from the satellite to produce land cover maps (e.g. taking elevation readings along with intensities from different types of electromagnetic radiation such as infrared, visible light, microwaves etc). This combination can be a cost-effective way of producing land cover maps as data from satellite imagery is becoming more available with open-source projects [36], [37], [38], [39] and the automated nature of MLTs helps in cutting the cost associated with creating models.

Within any mapping made through modelling, there is bound to be some degree of erroneous cases (as defined by disagreeing with a ground-truth assessment). These erroneous cases add a degree of uncertainty to how a land cover map would appear if it were to be made through a full census. A widely recommended practice in land cover mapping applications is to account for erroneous cases with UQ constructed using a sample of ground-truth assessments to act as reference data [40], [41], [42], [43]. However, the size of these samples and locations that can be visited can be limited by cost and practical restrictions (e.g. one may not be able to physically survey some areas within the mapping space). Such sampling limitations can ultimately lead to a high degree of uncertainty as estimates are left imprecise due to a lack of relevant ground-truth data. This is especially true when estimating the prevalence of rare or heavily clustered land cover types [44].

The purpose of the land cover mapping applications in this thesis is to act as case studies to test how adaptive sampling may help in collecting ground-truth data so that uncertainty can be quantified and reduced efficiently. Figure 1.3 illustrates the main components of the land cover mapping problems.



1. Satellite imagery and machine learning offer a cheaper and easier way of producing land cover maps.

2. Ground truth data is often needed to quantify uncertainty in these maps. This ground truth data may involve physical surveys, manual categorisations using higher resolution imagery, taking samples for lab testing etc.

3. However, collecting ground truth data is often expensive and subject to other restrictions (e.g., inaccessible areas). This can leave estimates with a high degree of uncertainty.

4. This thesis considers how ground truth sites can be selected so that uncertainty in the landcover maps is quantified and reduced efficiently through an adaptive sampling framework.

5. Using the lessons from landcover mapping applications, the thesis will then reflect on how the adaptive sampling framework can apply to other domains

Figure 1.3. *A summary of how the challenges in collecting ground-truth data impact uncertainty in land cover mappings and how the thesis aims to address this through adaptive sampling.*

From a more general perspective, the reason for using land cover mapping applications as case studies is that the issues mentioned here are representative of the challenges faced when trying to quantify and reduce uncertainty in MLTs (see Figure 1.4 for an illustrative overview). In this more general context, the land cover maps built with satellite imagery and MLTs are a substitute for classifiers built with MLTs and ground-truth assessments are a stand-in for manual categorisations. The problems surrounding the level of uncertainty from sampling limitations for ground-truth data are specific cases where there are restrictions in reference sampling. Hence, if one can develop an adaptive approach to reference sampling for land cover mappings, these methods should generalise to the wider problem of designing reference samples that efficiently manage the uncertainty in classifiers built using MLTs.

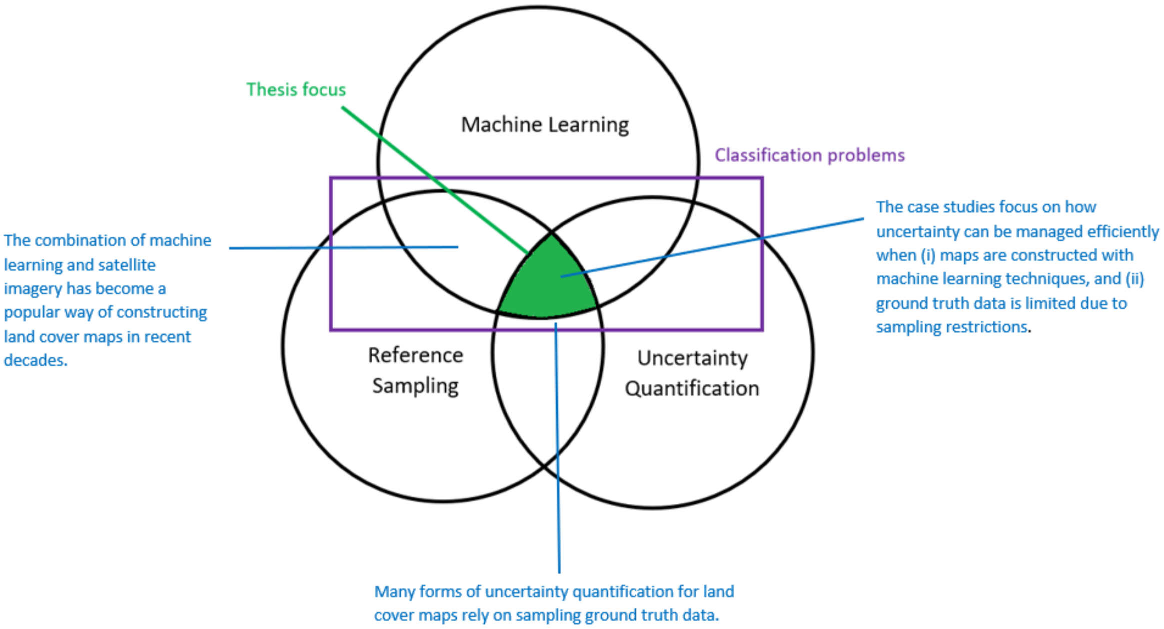

Thesis focus
Machine Learning
Classification problems
The combination of machine learning and satellite imagery has become a popular way of constructing land cover maps in recent decades.
The case studies focus on how uncertainty can be managed efficiently when (i) maps are constructed with machine learning techniques, and (ii) ground truth data is limited due to sampling restrictions.
Reference Sampling
Uncertainty Quantification
Many forms of uncertainty quantification for land cover maps rely on sampling ground truth data.

Figure 1.4 *An overview of how the land cover mapping applications fit in with the wider focus of this thesis.*

1.3. Programme of research

1.3.1. Aims and objectives

This thesis aims to investigate how one can better manage trade-offs between sampling restrictions and uncertainty in machine learning through an adaptive approach to sampling that is agnostic to the type of classifier and UQ. Here, adaptive sampling refers to a form of reference sampling whereby reference data are collected iteratively and uses the previous reference data to inform the design of future samples. Under adaptive sampling, the sample design of each iteration is free to vary in terms of size and which members of the population are targeted so that uncertainty can be managed efficiently.

The aim of this thesis is motivated by the need to create a cost-effective way of managing uncertainty in classifiers built using machine learning techniques. The motivation for an approach that is both classifier and UQ-agnostic comes from the problem that many machine learning classifiers are black-box in nature and there are often multiple viable MLTs and approaches to UQ based on different philosophies or assumptions one is willing to accept. Hence, by keeping any adaptive sampling practices agnostic to these choices, the work in this thesis has a better chance of being relevant to a broader range of problems.

Specifically, this thesis focuses on the following objectives:

A. Develop a framework for adaptive sampling that allows users to efficiently manage uncertainty in classifiers built with machine learning techniques. This framework should allow users to leverage the information contained in an initial reference sample to make informed and more cost-effective choices related to further sample designs when quantifying uncertainty under classifiers built using machine learning techniques.

B. Evaluate the proposed framework using a series of land cover mapping applications.

C. Provide recommendations on how the proposed framework could be further developed to address any unresolved weaknesses found in the evaluation stages.

1.3.2 Approach

This thesis constructs and evaluates a framework for adaptive sampling through an iterative process of reflection and refinement using a series of case studies involving land cover mapping problems. This is done as a means of developing methods and gaining insights into how adaptive sampling can be used to manage uncertainty in classifiers efficiently when they have been built using MLTs.

More specifically, this thesis uses two land cover mapping case studies that consider the problem of generating ground-truth sample designs to efficiently manage uncertainty within the following contexts:

(i) UQ under discrete classification maps.

(ii) UQ under fuzzy classification maps.

(iii) Sample design when the total size is limited by cost restrictions.

(iv) Sample design when the cost of ground-truth sampling varies across a region.

Figure 1.5 displays which areas each case study covers.

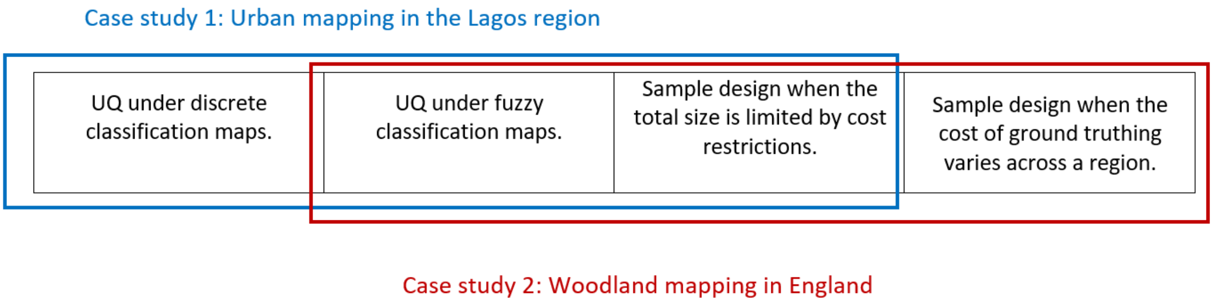
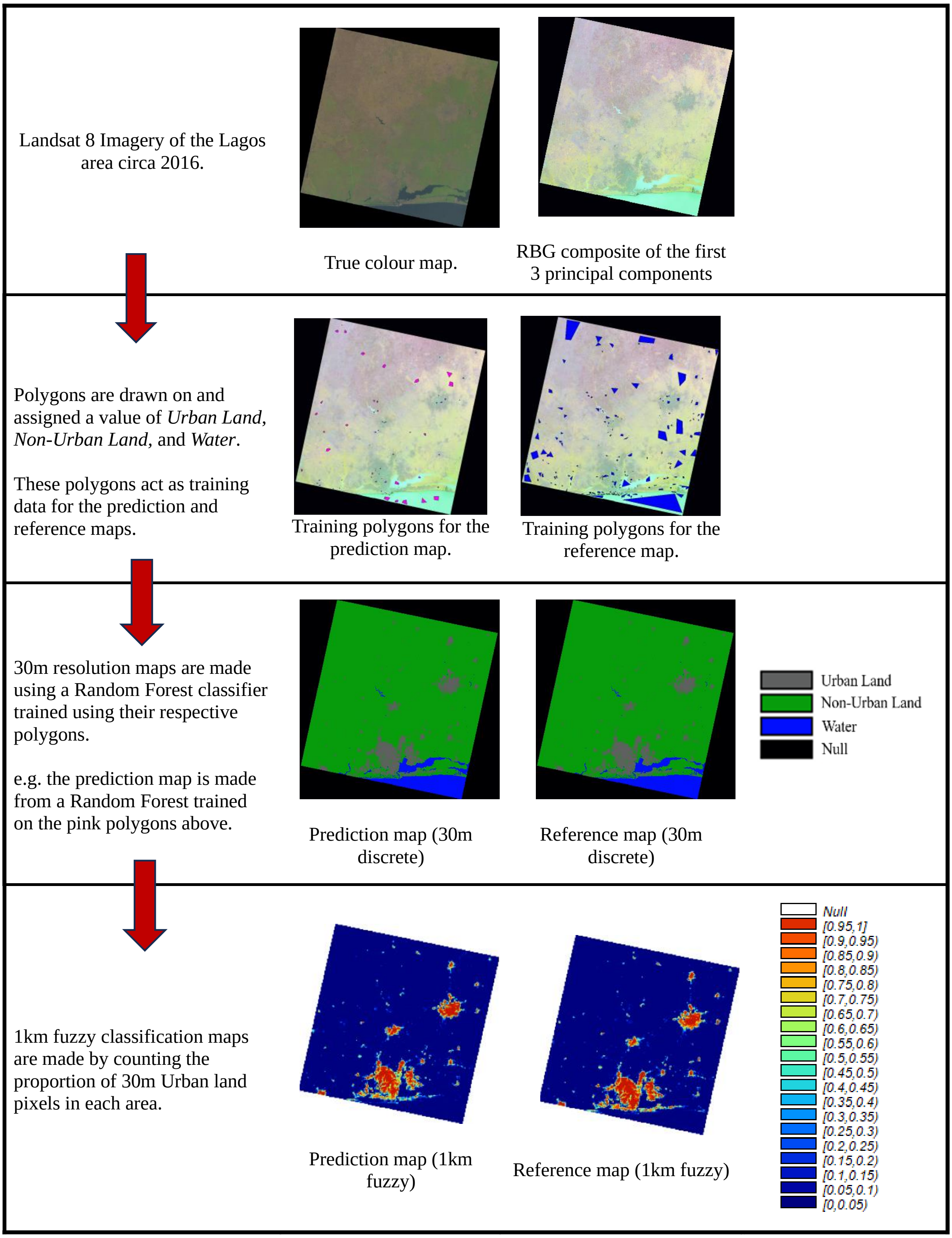

Case study 1: Urban mapping in the Lagos region
UQ under discrete classification maps.
UQ under fuzzy classification maps.
Sample design when the total size is limited by cost restrictions.
Sample design when the cost of ground truthing varies across a region.
Case study 2: Woodland mapping in England

Figure 1.5. *An overview of how the two case studies in this thesis cover the different classification scenarios.*

Note, these four situations are not an exhaustive list of scenarios. Rather, it reflects the types of problems practitioners often face. There are other issues and contexts to consider (Section 7.3 discusses this further) but through prioritisation, the thesis focuses on these four contexts.

Case study 1: Urban mapping in the Lagos region

The first case study involves an urban mapping of Lagos and the surrounding areas from early 2016. In this case study, two maps are provided, a prediction map and a reference map, both made using Landsat 8 imagery [45] and Random Forest classifiers [46]. Both maps classify pixels at a 30m resolution into one of three discrete categories: urban land, non-urban land and water. From this, two urban extent maps are generated at a 1km resolution, which represents the degree of urbanisation based on the proportion of urban 30m pixels that fall within each 1km square for each map. i.e., the predicted urban extent map is a set of 1km squares, with each square being assigned a value of 0 to 1 based on the proportion of 30m pixels in the same area that was classified as urban in the prediction map. This provides the case study with both discrete (30m) and fuzzy (1km) classification problems.



Landsat 8 Imagery of the Lagos area circa 2016.

Polygons are drawn on and assigned a value of *Urban Land*, *Non-Urban Land*, and *Water*.

These polygons act as training data for the prediction and reference maps.

30m resolution maps are made using a Random Forest classifier trained using their respective polygons.

e.g. the prediction map is made from a Random Forest trained on the pink polygons above.

1km fuzzy classification maps are made by counting the proportion of 30m Urban land pixels in each area.

***Figure 1.6.** A summary of how the prediction and reference maps are defined for the 30m and 1km maps.*

This case study is chosen to act as a bounded case study where initial ideas and concepts can be cheaply explored in early iterations in the context of both discrete and fuzzy classification problems, as one has access to both the prediction and reference maps. From a visual inspection of the maps in Figure 1.6, one can see that for a large proportion of the mapping area, the prediction and reference maps will agree, but there may be disagreements in areas near small towns or the outer region of cities. Hence, there is a suspicion before one starts that any optimal design for reference will want to target these areas more heavily.

The role of this case study is to give a space where one can investigate how well methods might learn to target these areas if one pretends that the full reference map is not available. From this, one can bring forward a shortlist of the better-performing methods for the second case study.

Another important property of this case study is the scale of the mapping area. One of the advantages of MLTs is that they can deal with large volumes of data (especially once fitted). Hence, if any method of generating sample designs for UQ is going to be suitable for applications involving MLTs, such methods will need to be capable of dealing with large volumes of data. In this case study, the 30m resolution maps consist of approximately 40 million pixels and the 1km resolution maps involve 36 thousand areas to consider, which provides a sufficient volume of data when evaluating the computational demand of any proposed methods.

Case study 2: Woodland mapping in England

The second case study involves quantifying the uncertainty for woodland mapping in England when the cost of collecting ground-truth data are high and the ability to collect data varies across the region due to travel restrictions motivated by COVID-19 regulations in 2020. In this scenario, one is faced with the problem of trying to generate a sample design that best manages the trade-offs between the degree of uncertainty in predictions; the costs associated with sending experts to perform physical ground visitations; and the additional COVID-19 travel restrictions that imply a strong preference to avoid sampling areas that are far from the locations of the experts.

Along with the predicted woodland map, there is also a propensity map that represents the preference for physically visiting some areas over others due to travel restrictions brought about by the COVID-19 virus (Figure 1.7B). Essentially, this map illustrates the preference that experts physically visit areas that are close to their home locations, which allows them to avoid overnight stays.

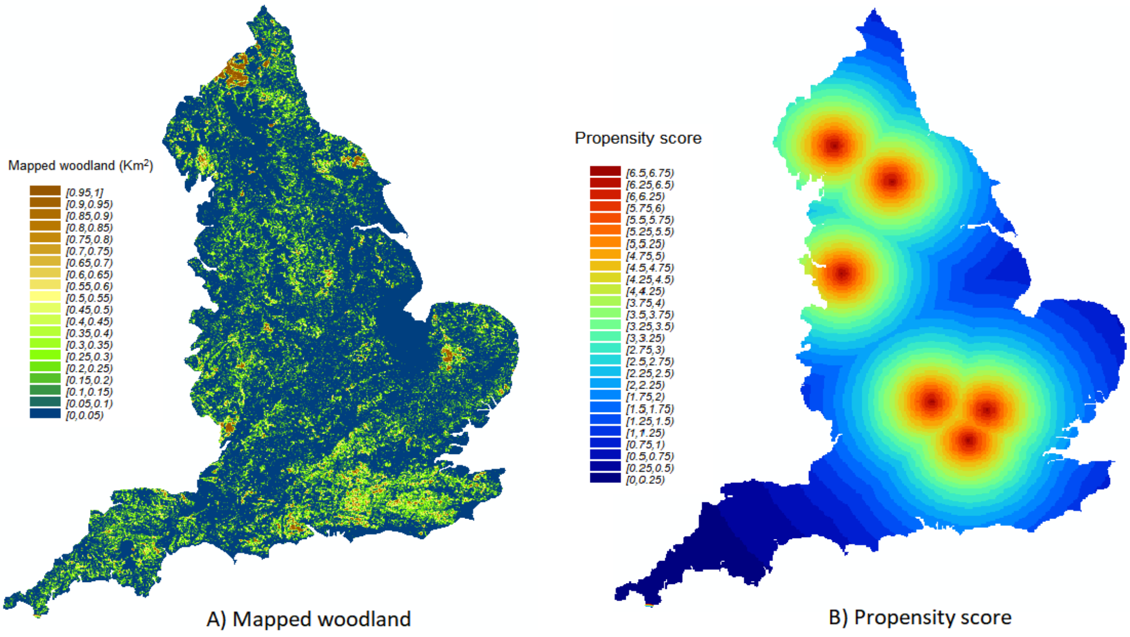


Figure 1.7. *(Left, A) Woodland mapping generated for England from* the *2015 UK land cover map. (Right, B) A mapping of the propensity scores based on the distance from where surveyors are located.*

The motivation behind the choice of this case study is to act as a secondary stage for evaluating and refining methods shortlisted in the first case study with a more challenging scenario. It maintains some of the important properties in the first case study as (i) there are approximately 130,000 1km squares in the mapping area which provide sufficient volume of data to be relevant to MLT applications; (ii) there is an expectation that misclassifications will be uncommon and clustered, which provides a suitable problem when evaluating targeted sampling practices.

However, it differs from the first case study by providing a more realistic scenario that better reflects the challenges related to UQ and reference sampling one is likely to face. Namely, one will not have the benefit of a full set of reference data at the beginning and standard sampling designs such as simple random sampling may not be viable in practice.

Additionally, the specific mechanics behind the propensity score will not be used in this case study. Hence, many of the results from this case study related to propensity scoring are likely to be relevant to classification problems in other domains, providing that a user can quantify their preference for sampling from different members of a population beforehand.

1.3.3 Overview of thesis

This thesis introduces an approach for managing the costs associated with quantifying and reducing uncertainty in classification problems by considering how one may collect the necessary reference data more efficiently. More specifically, the thesis is structured as follows:

- Chapter 2 provides an extensive literature review on the topics of uncertainty quantification, reference sampling, and machine learning and examines how the current literature relates to managing uncertainty efficiently in machine learning classifiers.
- Chapter 3 sets the evaluation criteria, proposes a framework for adaptive sampling, and populates said framework with a series of methods. Here, the framework represents adaptive sampling as an abstract process with four key stages and the methods are specific practices available to users that target these different stages.
- Chapters 4 and 5 use the Lagos and England woodland case studies respectively to evaluate the framework and methods introduced in Chapter 3.
- Chapter 6 reflects on case studies in Chapters 4 and 5 to evaluate the framework against the criteria set out in Chapter 3 at a more general level before discussing several important items emanating from this work.
- Chapter 7 summarises the work of the thesis and documents important areas of future work.

Chapter 2. Literature Review

When considering the problem of managing uncertainty efficiently in classifiers built using machine learning techniques, one is considering an intersection of three topics: uncertainty quantification (UQ), reference sampling and machine learning classifiers. The purpose of this chapter is to review the literature surrounding these three topics within this context. The structure of the chapter is as follows (see Figure 2.1 for an illustration): Section 2.1 reviews two key concepts behind quantifying uncertainty in classification problems: the first is design and modelling dependencies (2.1.1). The second is the choice between frequentist and Bayesian inference (2.1.2). Section 2.2 reviews the approaches one can take to reference sampling and how such methods interact with quantifying and efficiently managing uncertainty. Section 2.3 reviews the state of the art for classification algorithms made using MLTs. Finally, Section 2.4 reviews the literature for methods that look at reference sampling and UQ under MLTs, looking specifically at the intersection between these areas.

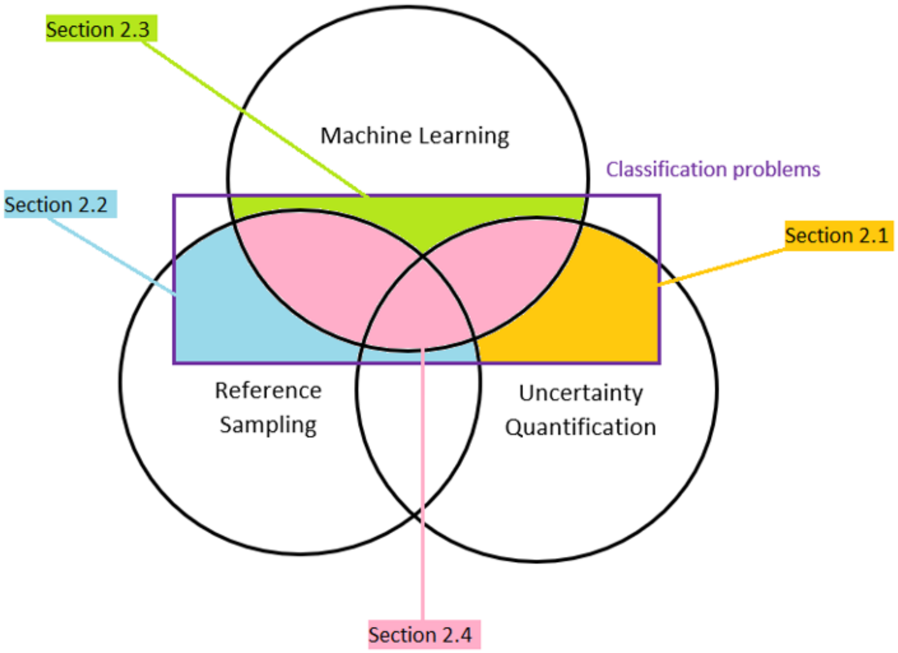


Figure 2.1 *A visual representation of the structure of Chapter 2 and how it relates to the focus of the thesis.*

2.1 Uncertainty quantification

Uncertainty quantification aims to provide a numerical representation of uncertainty using available evidence and reasoning. However, many subjective choices need to be made when choosing how to quantify uncertainty which makes the topic of developing efficient sample designs more difficult, as there is no guarantee that efficient sample designs under one method of UQ will be efficient under another method of UQ. In other words, the fact that there are subjective choices in how one quantifies uncertainty in the first place makes the idea of trying to create efficient sample designs more difficult as there are simply more scenarios to consider. Hence, before focusing on how to sample efficiently, it is important to understand how methods of UQ are chosen.

When choosing a method of UQ, there are two major considerations that this section will cover. The first is to what extent one should rely on sample design or modelling assumptions. The second choice is whether to base UQ on a frequentist or Bayesian perspective.

2.1.1 Design and modelling dependencies

Quantifying uncertainty is more than a simple calculation using reference data. Typically, there needs to be some additional assumptions placed as to how the reference data were collected (i.e. the design) or contextual information (modelling) before methods of UQ can be considered legitimate.

Traditionally, the reliance a method of UQ places on the sample design and modelling assumptions has been framed in terms of design-based inference and model-based inference. Here, there is a well-established literature discussing the pros and cons of each approach, both at a general level [47], [48], [49], [50] and across many domains including, sociology [51], forestry monitoring [52], [53], soils monitoring [54], neurology [55], and land use/land cover mappings [47], [52], [56].

In more recent years though, the discussion has moved away from the binary choice of model and design-based inference in UQ and more towards a discussion on how methods of UQ have different degrees of reliance on design and modelling assumptions [43], [57], [58]. These more fluid perspectives are useful for differentiating between different forms of UQ including model-assisted estimation [59] and hybrid estimation [43] (see Figure 2.2 for an illustration).

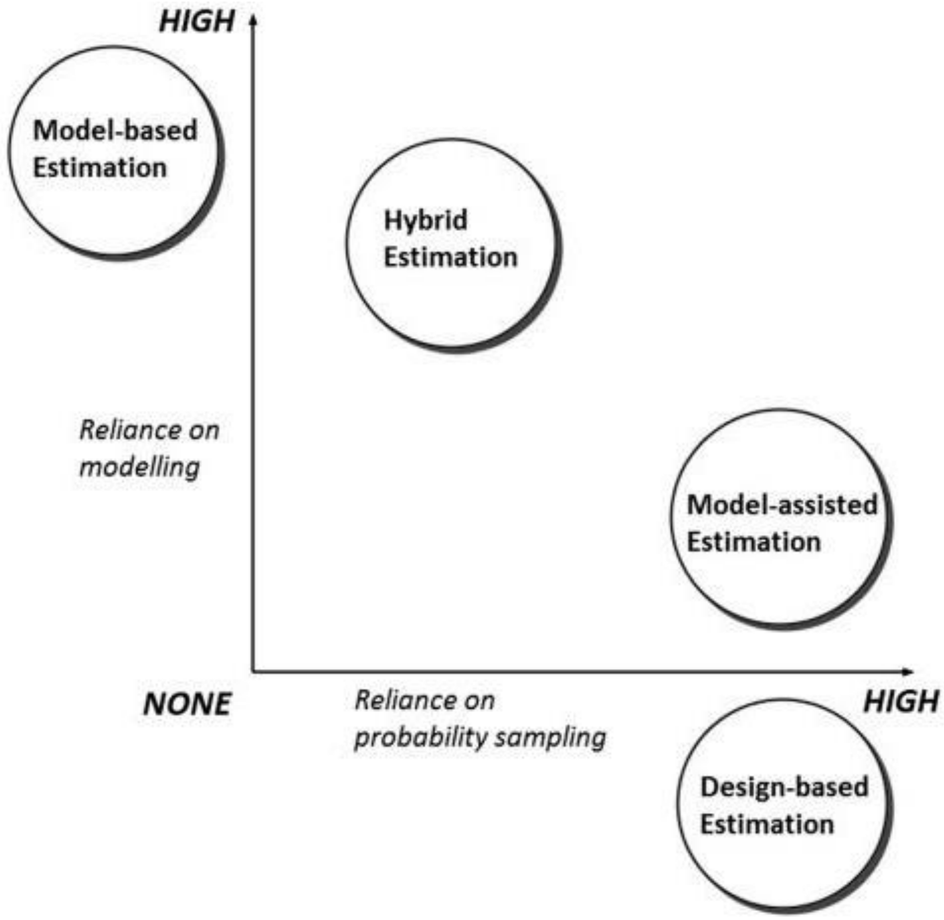


Figure 2.2. *An overview of how some methods of uncertainty quantification have different reliance on modelling and design assumptions. Original image from Ståhl et al* [43].

Because of this extra utility in a more fluid view of model and design-based estimation, this thesis will from now on adopt language that better reflects varying degrees of reliance on modelling and design assumptions in UQ. More specifically, the *design dependencies* for a method UQ refer to aspects that rely on the design of the sample. Likewise, the *modelling dependencies* refer to aspects that rely on additional modelling assumptions. A method of UQ with many (or few but severe) dependencies of one type of inference is said to be highly reliant on that type of inference.

For this thesis, several key themes stand out from the literature when discussing the advantages and drawbacks of different levels of modelling and design dependencies:

- As a method of UQ becomes more design-dependent, its validity depends more on being able to (i) implement strict probabilistic sample designs (see Section 2.2.1 for further details on probability sampling) and (ii), assign an explicit value to each member of the population that indicates the probability that said member will be included within such a design (these probabilities are often referred to as inclusion probabilities). In addition, when inclusion probabilities are close to or exactly zero because of sampling restrictions (e.g. one is unable to survey a particular subset of a population), one may see a *decrease* in the precision of estimates or force a user to exclude subpopulations in any analysis, as estimates may not be well-defined when inclusion probabilities are exactly 0.

- Purely design-based approaches are only suitable for population-level estimates (e.g. overall accuracy, population means, etc.). Hence, if one wishes to quantify uncertainty for estimates involving individual cases or small subsets of a population, some degree of modelling dependency will be necessary.

- Modelling can provide a way of increasing the precision of estimates under limited data by taking advantage of correlations between auxiliary variables and target variables. The exact balance of design and modelling dependencies can vary, with model-assisted estimators relying more heavily on design dependencies and less on modelling assumptions than, say, a purely model-based approach. However, one must be cautious when making a direct comparison between the precision of estimates with these approaches, as they will often rely on different modelling assumptions and sources of ontological uncertainty.

Data-driven vs process-driven modelling.

When making decisions related to design and modelling dependencies in UQ, it is useful to distinguish between data-driven modelling and process-driven modelling. Hunter et al. [60] describe process-driven modelling with

“*Process-driven models are developed from the known physical process(es) in a system, which are represented mathematically.”*

In contrast, Hesamia *et al*. [61] use the following to describe data-driven models

"Regarding data-driven modeling, data are analyzed in the system for investigating the relation with the system state variables without considering the physical behavior of the system."

Another way of viewing this difference is that process and data-driven modelling aims to quantify a physical process with relevant variables and use reference data to calibrate parameters within these models. On the other hand, data-driven modelling aims to form some sort of generalised structure for the relationship between variables. When the entirety of the data from a population is unavailable, such relationships must be estimated with a sample. Whilst MLTs are not strictly required for data-driven modelling, their use has made the distinction between process-driven and data-driven modelling more apparent in many domains including mechanical engineering [62], plant-wide industrial processes, [63], clinical drug development [64] and fault diagnosis in nuclear power systems [65].

In terms of UQ and the relation to modelling and design dependencies, the differences between process-driven and data-driven modelling can be seen as specific cases of the idea presented in Figure 2.2 (see Figure 2.3), with more modelling dependencies placed on process-driven modelling (from the fact that one will often need to predefine any causal relationships between variables) and more design dependencies in data-driven modelling as causal relationships between variables are not assumed.

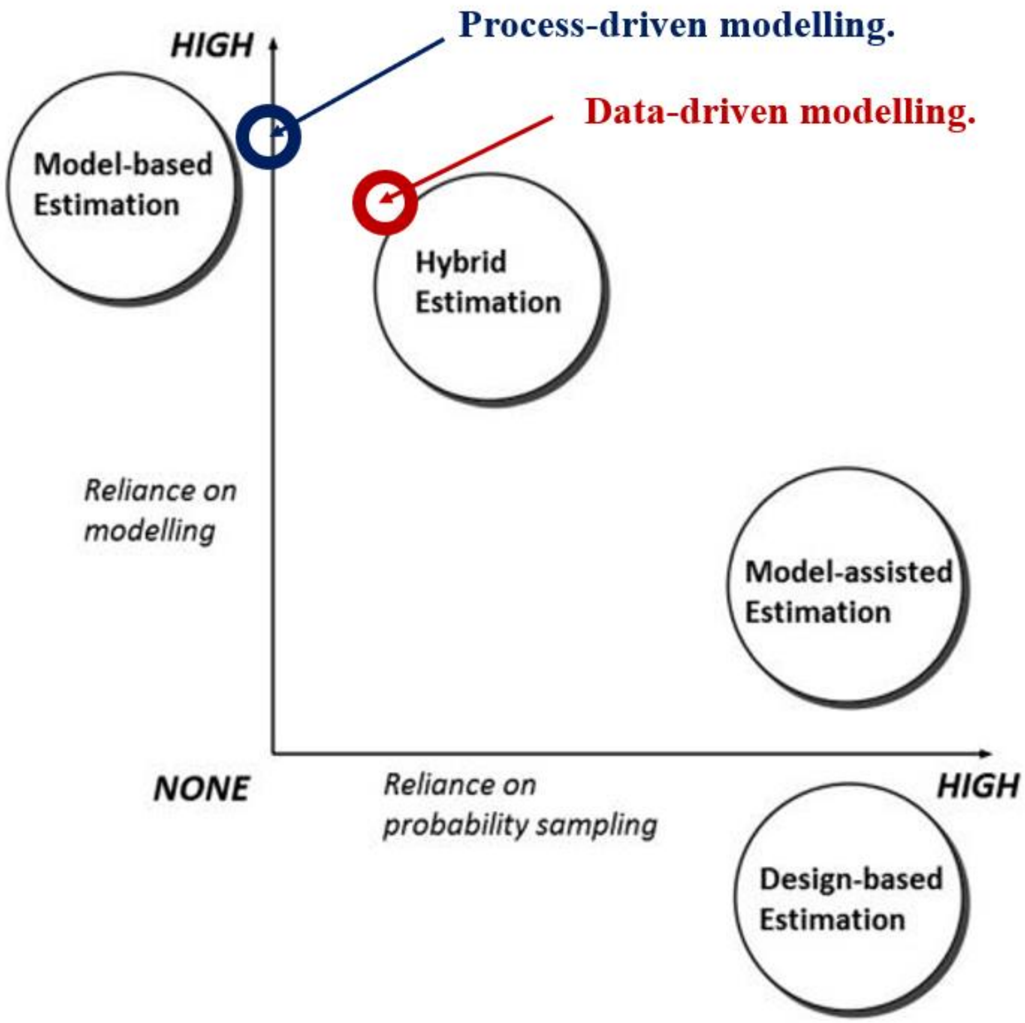


Figure 2.3*. A representation of how process-driven and data-driven modelling relates to reliance on design and modelling assumptions.*

The motivation for highlighting the distinction between process-driven and data-driven modelling in particular is that, for many MLTs, it is not easy to quantify uncertainty through process modelling due to a lack of interpretability and explainability [66], [67]. Consequently, the use of MLT classifiers will typically come at the cost of forgoing process-driven modelling in UQ, which in turn will inevitably mean a reliance on either data-driven modelling or some form of UQ with a higher level of design dependency.

2.1.2 Frequentist and Bayesian inference

At a high level, a method of UQ aims to represent the perceived likeness of an unknown value or event based on currently available information. Typically, this involves representing uncertainty in terms of probabilistic statements, to which there are two major interpretations of probability, frequentist and Bayesian [68], [69].

From a frequentist interpretation, probabilities are based on proportions related to a large number of events. For example, if a fair coin is flipped many times, one would expect around half of the flips to land on heads, and this would be the basis of assigning a statement of the form “each flip on the coin has a fifty per cent chance of landing on heads”. For uncertainty quantification, this frequentist idea of probability is called upon when producing confidence intervals. Neyman [70] defines confidence intervals with:

An X% confidence interval for a parameter θ is an interval (L, U) generated by a procedure that in repeated sampling has an X% probability of containing the true value of θ, for all possible values of θ.

Using frequentist approaches in discrete classification problems is well-established for many common estimates [71]. For example, when estimating performance metrics in discrete classifiers from a sample (e.g. overall accuracy, sensitivity, specificity etc.), this can often be viewed as estimating binomial proportions. From this, there are many well-known approaches for producing confidence intervals [72]. When deriving a closed form for the distribution of an estimate is more difficult, one can rely on simulation-based methods such as bootstrapping [73]. Such methods are commonly used when estimating metrics such as the area under the ROC curve [74], [75].

For fuzzy classification problems, there is a wide variety of frequentist-based UQ methods centred on regression analysis [76] and model-assisted estimation [77] that one can draw upon.

With Bayesian inference [78], [79], uncertainty in θ given a sampled set of data D is represented as a probability distribution, $\pi(\theta|D)$, by using Bayes theorem:

$$\pi(\theta|D) = \frac{\pi(D|\theta)\pi(\theta)}{\pi(D)},$$

where $\pi(D|\theta)$ is the likelihood function (often defined using the sample design and other assumptions), $\pi(\theta)$ denotes the prior distribution, and $\pi(D)$ is the marginal likelihood. The prior distribution reflects belief in the likely values of θ before observing the data, D. The choice of prior distribution may be influenced by many factors such as previous studies; the context of the problem (e.g. knowing values are bounded by definition); the subjective belief of the user; a desire for the posterior distribution to be primarily influenced by the observed data (this is commonly referred to as using a vague or non-informative prior distribution [80], [81]), or for mathematically desirable properties that allow for closed form posterior distributions (e.g. conjugate priors [82], [83], [84]). In general, when the choice of the prior distribution is not clear, a sensitivity analysis is recommended [85]. The calculation of $\pi(D)$ is often avoidable in practice as it only acts as a normalising constant.

With any posterior distribution, one has the additional option of a Bayesian analogue to confidence intervals with credible intervals [86]. The difference between Bayesian credible intervals and frequentist confidence intervals is that credible intervals are a measure of uncertainty that captures the most likely values for θ based on the posterior distribution. For example, an equal-tailed $X\%$ credible interval for θ given D would be the $\frac{X}{2}$-th and $\left(100 - \frac{X}{2}\right)$-th percentiles of the posterior distribution $\pi(\theta|D)$.

There has been substantial debate over the benefits and drawbacks of frequentist and Bayesian inference at a general level [68], [69], [87]. When considering the method of UQ under restricted sampling designs, these arguments can be usefully broken down into the following headings: sources of ontological uncertainty; the ability to deal with design-based approaches; the ease of applying simulation-based methods.

Sources of ontological uncertainty

With frequentist methods, a key source of ontological uncertainty comes from the potential difference between the nominal coverage and the stated coverage. The nominal coverage is the true proportion of confidence intervals (or regions when considering higher dimensional estimates) that would contain any unknown value should the sample design be repeated a large number of times. In an ideal situation, the nominal coverage is exactly (or at least close to) the stated level of confidence.

In some well-structured situations, this can be guaranteed by pivoting key statistics [88]. Briefly, such methods begin with a statistic for θ derived from the sampled data. When this statistic satisfies a number of key properties, one can then use the distribution of this statistic to reverse engineer a confidence interval for a θ at a specified level of confidence with the guarantee that the nominal coverage will match this.

Mismatches between stated confidences and nominal coverage occur in two common scenarios. The first scenario is when the distribution of a statistic for θ is discrete. As a quick description, this is because, when a distribution of a statistic is discrete, the cumulative distribution function (CDF) for that statistic will not be continuous, which makes it impossible to invert this CDF at the points that would ensure an exact level of coverage. One well-known example of this is when estimating binomial proportions with Clopper-Pearson intervals [89]. Here, using the exact distribution can be overly cautious in some situations, i.e. the nominal coverage is so much greater than the stated level of confidence, that the confidence intervals are unnecessarily wide [90]. This can be a substantial drawback when sample sizes are limited and may motivate one to look at alternative statistics that sacrifice exact coverage for narrower confidence intervals with coverages that are close enough approximations to their stated level of confidence. In the case of binomial proportions, this can be seen in Agresti–Coull intervals [91] or using methods based on Bayes theorem [92].

The second scenario is when one is relying on approximate distributions of statistics. Such approximations are useful when: (i) deriving the exact distribution for a statistic is overly cumbersome or analytically impossible; (ii) one is wishing to avoid other sources of ontological uncertainty by making assumptions necessary for exact methods. Two well-known examples of these methods are applications involving the central limit theorem [93], and bootstrapping methods [94], [95]. The central limit theorem and bootstrapping are highly generalisable methods based on asymptotic theory that offer a means of approximating distributions of key statistics under sufficiently large sample sizes without making strong assumptions about the distribution of a population.

In either case, once one begins deviating from using the exact distribution of statistics (which is often necessary in practice), there is an underlying source of ontological uncertainty regarding whether any stated level of confidence is sufficiently close to the nominal coverage. In the case of estimating a binomial proportion from a simple random sample, previous work has been able to demonstrate that differences between the nominal coverage and stated confidence for approximate methods are sensitive to the true proportion value, the sample size, and the stated level of confidence [91]. Because of the relative simplicity of these cases, one can empirically assess these relationships. However, analysis of this kind does not extend well to more sophisticated scenarios, as visualising results across multiple dimensions can quickly become difficult.

Two noteworthy situations where it can become difficult to formulate exact distributions for statistics are multilevel modelling [96] and adaptive sampling. To be clear, there are frequentist methods capable of providing confidence intervals under multilevel modelling adaptive sampling. In multilevel modelling, one can use asymptotic properties of maximum likelihood estimates [97] or bootstrapping [98] to construct confidence intervals. Likewise, it is possible to construct confidence intervals with sufficient coverage under adaptive sampling when the decision-making processes involved can be explicitly quantified. Examples of such methods can be found in clinical trials [99], [100], [101], [102], [103] and estimating insect populations [104]. The point to emphasize here is that should one need to rely on approximate methods in these situations, confirming nominal coverage empirically is difficult due to an increase in the number of factors involved. With multilevel modelling, one must consider nominal coverage across all true values of all the estimates involved. With adaptive sampling, any factor in the decision-making processes (e.g. sample size within batches, stopping rules, targeted sampling etc.) must be considered in any analysis.

The net result of this is that, when using frequentist methods to quantify uncertainty under limited sample sizes, one is forced to either (i) restrict sample designs to simple situations so that one can use methods that can ensure a reasonable match between coverage (either analytically or empirically), which may, in turn, limit the efficiency of any uncertainty reduction, or (ii) rely on methods that have a realistic possibility of producing invalid measures of uncertainty, with no easy way to confirm the validity of said methods.

In comparison, under Bayesian inference, the posterior distribution is a purely logical statement based on the observed data, prior distribution, and likelihood function. In other words, if one believes that the likelihood function and prior distribution are appropriate, then the posterior distribution is an appropriate quantification of uncertainty in θ after observing D. This gives Bayesian techniques an advantage over frequentist methods under limited sample sizes, as it allows one to bypass any concerns regarding stated confidence and nominal coverage. The potential drawback for this is that sources of ontological uncertainty are now more heavily placed on the appropriateness of the prior distribution and the assumptions necessary to formulate likelihood functions.

Another key difference with Bayesian inference is that, because uncertainty in θ is represented as a probability distribution, one can apply well-known results from probability theory by treating the posterior for θ, $\theta|D$, like any other random variable. In particular, it allows one to make use of three techniques, marginalisation [68], [79], Monte Carlo integration [105], and Markov Chain Monte Carlo (MCMC) methods [106], [107]. These techniques are powerful tools when dealing with more sophisticated situations such as multilevel modelling. For example, suppose one is interested in some function of $\theta, g(\theta)$. Through marginalisation, one can express the posterior distribution for g as:

$$\pi(g(\theta)|D) = \int \pi(g(\theta)|\theta, D)\pi(\theta|D)d\theta.$$

When no analytical solution for $\pi(g(\theta)|D)$ is available, one may use Monte Carlo integration to approximate $\pi(g(\theta)|D)$ if one can sample from the posterior distribution, $\pi(\theta|D)$, and apply g under fixed values of θ. The same principle can be applied multiple times to include more advanced cases of multilevel modelling. MCMC methods offer a way of approximating the posterior distribution, $\pi(\theta|D)$, if analytical solutions for $\pi(\theta|D)$ are not available. The combination of Monte Carlo methods and marginalisation is a key advantage of Bayesian inference, as it allows one to move to more sophisticated scenarios such as multilevel modelling without needing to rely on asymptotic theory or difficult-to-obtain analytical solutions [108]. It is important to stress here that the precision of Monte Carlo approximations is governed by the number of simulations and not the size of any reference sample. This means that, given enough simulations and computational resources, one can reach an arbitrary degree of precision for a posterior distribution with Monte Carlo methods. In contrast, frequentist methods rely on asymptotic results that are approximations with respect to the amount of reference data available. Since generating simulations is generally far easier than collecting more reference data, concerns over the suitability of approximations are unlikely to be as prevalent here.

Applications of multilevel modelling do not need to be particularly sophisticated for this advantage of Bayesian inference to be useful in classification problems. One example of this is when comparing the performance of multiple classification algorithms simultaneously [109]. With frequentist approaches though, the shortcomings of significance testing make similar analysis much more difficult.

Another situation relevant to classification problems is when dealing with stratified sampling. One example of this is when estimating deforestation under stratified sampling in land change mapping [110]. In this example, the total area of deforestation, $A = \sum_{i=1}^{n} W_i p_i$, is a weighted sum based on the n strata, with stratum sizes, W_i, and the proportion of deforested area in each stratum, p_i. Such a case can readily be handled with Bayesian inference by generating a posterior distribution for each p_i and approximating a posterior distribution for A with simulation-based methods. In comparison, many of the frequentist methods considered are either not suited for stratified sampling or are called into question because of a combination of modest sample sizes within some strata and that some highly weighted strata were expected to have almost no deforestation (i.e. there were strata with a large W_i and p_i very close to, if not exactly equal to, 0).

Finally, the third advantage of Bayesian inference is the ease with which one can quantify uncertainty under iterative sampling. With Bayesian inference, a posterior distribution is the same regardless of whether the entire reference sample is viewed as a single batch or viewed as a series of subsamples that are updated sequentially [111]. This can be neatly summarised with the phrase *today's posterior is tomorrow's prior* [112] and be mathematically represented with the result that for two observed data sets D_1, D_2

$$\pi(\theta|D_1, D_2) \propto \pi(D_2|\theta)\pi(\theta|D_1).$$

The benefit of this property over previously mentioned frequentist approaches is that it allows one to treat batches of samples as a single sample and vice-versa, which greatly simplifies UQ under adaptive sampling. This property is not available in frequentist inference, as the latter must account for decision-making processes made during each iteration of sampling.

The ability to deal with design-dependent inference.

When deciding upon frequentist or Bayesian inference, it is important to consider how they interact with methods of UQ as design and modelling dependencies vary (see Section 2.1.1 for further details on design and modelling dependencies). For instance, many popular design-based approaches are based on frequentist inference, and there is not always a Bayesian equivalent to draw upon. One notable example of this is in model-assisted estimation [59]. Model-assisted estimation is popular in applications such as forestry monitoring when estimating large-scale quantities such as total deforestation [113]. Model-assisted estimation allows one to make use of auxiliary information to make precise estimates without needing to rely on formal modelling assumptions, which can easily become difficult to manage (e.g. accounting for structures regarding spatial correlations). Developing a Bayesian equivalent to model-assisted estimation is not as easy, as one will, at some point, need to impose further modelling assumptions to generate a suitable likelihood function. Depending on the circumstances, this advantage may be substantial enough to overshadow other advantages Bayesian inference may bring.

The ease of applying simulation-based methods

Simulation-based methods such as bootstrapping in frequentist inference and MCMC sampling in Bayesian are popular techniques in modern statistical inference that offer highly generalisable approaches for uncertainty quantification. Putting aside any sources of ontological uncertainty from the assumptions made or the philosophical differences for a moment, and focusing only on the application of each method, bootstrapping is arguably a much simpler method that is easier to implement in practice.

With bootstrapping, one can construct confidence intervals for an estimate by simply resampling from the observed data (with replacement) and fitting estimates based on these resamples. Hence, if one can calculate an estimate based on the observed data, there is little stopping a person from applying bootstrapping methods.

MCMC sampling methods such as random walk Metropolis-Hastings (RW-MH) allow one to sample from posterior distributions under almost any prior distribution and likelihood function [114]. With RW-MH though, one must carefully set the proposal distributions, striking a suitable balance between visiting areas where the posterior distribution is of higher density more often and exploring the space the posterior distribution occupies without getting stuck in small areas [115]. As this posterior distribution increases in dimensionality, fine-turning these proposal distributions becomes increasingly more difficult.

Specifically, on RW-MH, there has been substantial work on methods to propose such distributions [116], [117], [118], [119]. There has also been substantial work for approaches that aim to mitigate against problems involved in setting proposal distributions, if not avoid them altogether. This includes making use of hierarchical structures in modelling to break a problem of generating a high-dimensional posterior distribution into several lower-dimensional problems, and linking them together with RW-MH and Gibbs sampling [120], [121], [122], or using of likelihood-free approximations [123], [124] as a means of generating posterior distributions (note, the latter methods are sometimes referred to as approximate Bayesian computation). In some specific circumstances, it is possible to use bootstrapping to generate posterior distributions [125].

Whether these methods are useful to a specific problem will depend on the context of any situation. Furthermore, problems related to the efficiency in MCMC sampling may become less of an issue with the increasing availability of cloud computing services [126], [127].

However, the fact that bootstrapping is easier to apply without needing to rely on more advanced methods is an advantage that, for now, cannot be ignored. Depending on the context of the problem, this advantage alone may be enough for users to adopt frequentist inference in UQ, especially when the philosophical difference between frequentist and Bayesian inference is not a major concern.

2.1.3 Reflections

Careful consideration of how uncertainty is quantified is vital in any research that aims to investigate how uncertainty can be managed efficiently. Given that there are an infinite number of possible approaches to UQ, it is not possible to systematically review every method individually. However, it is possible (and useful) to consider the fundamental philosophies that underpin such methods. This section has considered design and modelling dependencies (2.1.1) along with frequentist and Bayesian inference (2.2.2).

Focusing on the choice between design and modelling dependencies first, there has been a substantial amount of work evaluating the pros and cons of design-based and model-based inference when quantifying uncertainty. More recent work has focused on blurring this dichotomy and allowing for hybrid approaches, which aim to strike the most appropriate balance between design requirements and modelling assumptions.

The debate between frequentist and Bayesian inference is also well-studied. Whilst Bayesian inference goes back to the original founding of Bayes theorem, the increase in computational capacity and the development of MCMC methods have led to a growing interest in Bayesian inference throughout the late 20th and early 21st centuries.

From the perspective of developing sample designs to manage uncertainty efficiently, this thesis draws the following key observations:

- In situations where sample sizes and designs are limited, methods of UQ based on fully design-based inference may not be efficient at managing uncertainty. This is especially true when sampling from some subsets of a target population is difficult or impossible.

- Discussions related to the appropriateness of different design and modelling dependencies are largely going to be domain-specific and context-specific. Furthermore, experts within a domain may even disagree on such issues. Hence, any framework for managing uncertainty efficiently should be as generalisable as possible and not rely on one specific set of design requirements or modelling assumptions.

- Bayesian inference offers several advantages over frequentist inference. These include the ability to (i) easily propagate uncertainty via simulation-based methods, (ii) naturally handle sequential sampling (iii) formally include prior knowledge when quantifying uncertainty.

- There are some niche situations where some frequentist methods of UQ may be preferred over Bayesian approaches. Two notable examples include using model-assisted estimators and methods using bootstrapping. However, these methods are based on asymptotic theory (along with many other popular frequentist-based methods). This can be a problem when sample sizes are limited, as it will not always be clear when (or if) such methods are valid.

2.2 Reference sampling

When considering how one can best balance different trade-offs between the costs of collecting reference data and uncertainty, how the reference data are collected (i.e. the sample design), and the types of reference data collected are naturally going to play an important role. This section reviews the current literature on how one may construct sample designs that can efficiently manage uncertainty. This thesis reviews such work across two complementary themes. Section 2.2.1 reviews sampling techniques from the perspective of probability vs non-probability sampling. Section 2.2.2 reviews methods for creating efficient sample designs.

2.2.1 Probability and non-probability sampling

A sample design is said to use *probability sampling* if members of the population are selected through a *known* probabilistic mechanism that can be expressed using a probability density function (pdf). In contrast, *non-probability* sample designs select members with methods that cannot be expressed using a pdf. The lack of a known pdf may be caused by one of two cases. The first case is when the sample design does not contain any stochastic components, meaning there is no underlying probabilistic mechanism. The second case is when there may well be a probabilistic mechanism governing which members are selected, but one is unable to describe this with a pdf due to a lack of understanding. Popular examples of probability and non-probability sampling are presented in Table 2.4.

Table 2.4. *An overview of popular examples of probability and non-probability sampling*

Probability sampling	**Description**
Simple random sampling	Each member of the population is selected independently and with an equal probability of being selected.
Stratified random sampling	The population is first partitioned into smaller groups known as strata. A simple random sampling is then conducted within each stratum, where each stratum is treated as an independent sub-population.
Systematic random sampling	Systematic random sampling begins by arranging the population within some ordered frame (e.g. alphabetical order). A member of the population is then selected through simple random sampling to act as a starting point. With this starting point, the remaining members of the population are then selected based on some periodic rule (e.g. every fifth member counting from the starting point).
Cluster and multistage sampling	The population is first arranged in groups called clusters. Following this, a set number of clusters are selected through a known probability sampling (e.g. simple random sampling or weighted according to cluster size). All members of the selected clusters are included in the sample. With multistage sampling, further sample designs are implemented within each cluster (e.g. simple random sampling, an additional clustered sampling etc.).
Non-probability sampling	**Description**
Quota sampling	The members of the population are selected to meet a specific objective based on some quota (e.g. a sample of 100 male and 100 female students).
Snowball sampling	First-level members are selected using some initial sample design. These first-level members are then used as a basis for selecting new members (e.g. using nearby members, first-level members recruiting new members etc.). This procedure is repeated iteratively to create a "snowballing" effect when sampling members. Snowball sampling may also be referred to as chain sampling or chain-referral sampling.
Convenience / purposive sampling.	The selection of members is based on some predetermined criteria. These criteria may include conveniences in sampling or other judgments made by the researchers (e.g. selecting extreme events, sampling experts for opinions etc.).
Self-Selection Sampling	Members of the population are given to the researchers via a voluntary process that is not controlled by the researcher.

The advantages and disadvantages of different types of probability sampling are well-studied [128], [129], [130]. The key takeaways in the context of developing efficient sample designs are:

Probability sampling is a requirement for design-dependent approaches of UQ.

As discussed in Section 2.1.1, some methods of UQ make use of design dependencies, meaning that they use probability designs as part of their fundamental basis. Hence, if one is wishing to employ design-dependent methods of UQ, some form of probability sampling is going to be necessary.

Non-probability sampling is often more convenient than probability sampling.

Non-probability sample designs are often a response to the practical difficulties found in probability sampling (or when there is a lack of control over how the sample is obtained). For example, quota sampling is useful when one is unable to obtain a probability sample but is trying to create a sample that is representative of the population being studied [129]. In such a case, quota sampling may be viewed as a non-probability analogue of stratified random sampling. Self-selection and snowball sampling can be useful when finding consenting members of a population is difficult [131], [132]. Purposive sampling may benefit classification problems that deal with rare events or multiple classes, as this can be a way of ensuring one has enough reference data from all categories.

UQ from non-probability sampling is possible but requires some degree of modelling assumptions.

In general, it can be difficult to use data from non-probability sampling in many methods of UQ, especially if the method of UQ relies on many design dependencies. However, it is possible to quantify uncertainty in estimates with such data with the addition of modelling dependencies.

One option is to use a fully model-based approach (see Section 2.1.1 for a discussion on the advantages and disadvantages of model-based approaches). Examples of using model-based approaches to quantify uncertainty under non-probability sampling can be found in soils monitoring [133], [134], forestry monitoring [135], [136], and market research using online surveys [137].

An alternative option for including non-probability sample designs is via propensity scoring [138], [139], [140]. Propensity scoring is based on the idea a non-probability sample design may be modelled as a probability sample under the right circumstances. Propensity scoring has been shown to be popular in clinical trials [141], [142], [143] as well as internet surveys [144], [145], [146], [147]. Propensity scoring has recently been considered in land cover mappings, where the use of volunteered reference data are becoming more common [148]**.**

Regardless of the specific modelling practices, the core premise of these approaches is that the problem of managing uncertainty efficiently can be addressed by making easily obtainable reference data suitable for UQ. In short, modelling and propensity scoring focuses on bringing cheaper designs into UQ as opposed to more traditional approaches that look to spend resources more carefully under expensive designs that are already suited for UQ.

The drawback of modelling and propensity scoring though is that there is inevitably going to be some form of ontological uncertainty from the additional modelling assumptions. Determining when this extra ontological uncertainty is worth the cost of any additional reference data is always going to be context-specific and prone to subjective choices.

Different forms of probability sampling provide different trade-offs between sampling convenience, efficiency, and ease of UQ.

Whilst all design-dependent approaches to UQ require probability sampling, not all probability sample designs are equally as efficient at reducing uncertainty. Here, there are three factors to consider: the ease of implementing the sample design in practice, how efficient the sample design is at reducing uncertainty, and the assumptions or conditions necessary to quantify UQ.

At a general level, simple random sampling lies at one extreme of this trade-off balance. With simple random sampling, UQ is typically the most straightforward, as there are few (if any) additional assumptions or further information required beyond those that already exist in the original model. The drawback of simple random sampling is that it can be difficult to implement in practice in many situations and may be inefficient in some applications. From this, other forms of probability sampling can be viewed as a trade-off, where extra sampling convenience or efficiency is achieved at the expense of additional assumptions in UQ.

Example 2.5 illustrates this idea of balancing trade-offs in the context of a land cover example. Even in this relatively constrained example, one can see how context-specific details such as sampling cost across an area, design costs being dependent on how far away the selected sites are from each other, and expectation of spatial autocorrelation are playing a substantial role in how one may choose a type of sample design. Sensitivity to details like these makes it difficult to say much about the pros and cons of sample designs at a general level. Instead, one would expect to conduct an analysis similar to example 2.5, yet context-specific, when dealing with new situations.

Example 2.5: Balancing trade-offs in probability sampling in land cover mapping applications

Suppose that one has constructed a land cover map and wishes to collect reference data using physical surveys to estimate the accuracy of a land cover map and the total area for different types of land cover. For the sake of simplicity assume that one has already decided upon a design-based estimator for UQ and needs to decide if the reference data should be collected under simple random sampling, stratified random sampling, systematic random sampling, or cluster sampling (single staged).

However, collecting reference data often involves physically visiting sites to conduct surveys, which can be expensive, especially if they are far away from each other or in hard-to-reach places. This in turn creates a need to carefully consider the efficiency of sample designs. In addition to this, there is a general expectation of spatial autocorrelation in land cover mapping applications [149], [150], [151]. This is because types of land cover typically cluster in areas and misclassifications are not usually uniformly distributed but tend to also appear in clusters (e.g. near border regions, specific classes misclassified etc.). With this contextual information, one can begin to evaluate the pros and cons of the different design options.

Under simple **random sampling**, one can use many well-known results to easily provide unbiased estimates for the variance of key statistics, making UQ straightforward. The drawbacks though are that firstly, the sampling costs may be unpredictable and potentially overly burdensome. This is because, under simple random sampling, any combination of members (under a fixed size) has an equal chance of being selected. Hence, if one is unfortunate enough, one may happen to draw a selection where many of the sites are far away from each other or contain an unusual number of hard-to-reach areas. In addition, simple random sampling may also be inefficient when spatial autocorrelation is present, as one needs to rely on chance alone to select enough sites from unique clusters.

Systematic random sampling based on spatial grid sampling can be an efficient method of sampling when positive spatial autocorrelation is present. The drawbacks though are that (i) the individual spaces will be far from each other (adding to sampling costs) (ii) providing unbiased variance estimates for key statistics can be difficult [152], and may require additional modelling assumptions to implement [153], [154], [155], (iii) Batch sampling is much harder under this kind of sampling, as providing unbiased variance estimates under multiple iterations is especially difficult when compared to the other design types here.

Under **stratified random sampling**, UQ is not too difficult as one can provide unbiased variance estimates that are relatively easy to calculate for many situations. One notable exception is when strata are homogenous, as variance estimates may be unreliable (and possibly undefined) when one does not observe enough positive and negative cases [110]. Note this is not a niche or non-trivial issue in land cover mappings though, as the presence of spatial autocorrelation and accurate mappings makes homogenous strata more likely. One way to avoid this problem would be to ensure enough positive and negative cases are observed across strata, but this may create a high barrier to entry for the total sample size, particularly if there are many homogenous strata.

Assuming variance estimates are reliable though, sample sizes for each stratum are free to vary according to convenience (increasing sampling convenience) and potentially optimised under a fixed sample size (increasing the overall efficiency), which can be extended to include different sampling costs within each stratum (see Section 2.2.2 for further details).

Cluster sampling may act as a means of keeping sampling costs down by selecting groups of sites that are close together. Unbiased variance estimates are available for UQ, even when cluster sizes are different and there the cluster has an unequal chance of selection [59] (increasing UQ convenience). However, the formulae for these estimates are noticeably more complicated when compared to simple random sampling.

At a general level cluster sampling can be inefficient [129] and becomes worse as clusters become more homogenous. When clusters are equally sized and randomly selected, this relationship can be quantified explicitly in terms of variance inflation factors and intra-cluster correlations [156]. This point is especially relevant to land cover mappings, as clusters are more likely to be homogenous because of spatial autocorrelation.

Here, there is an interesting dynamic when considering the efficiency of clustered sampling. Whis clustered sampling may be less efficient than other designs with equal sample sizes, it may be more efficient from a cost perspective as it is much easier to collect larger sample sizes. When exactly the benefits of cheaper sampling will outweigh the general inefficiency will naturally depend on the specific features of a problem (e.g. the cost of sampling against degrees of spatial autocorrelations).

With this analysis, one can begin to draw conclusions about the overall the pros and cons of different design types in land cover mapping applications. Figure 2.5 summarises how these design types compare against the ease of implementation, efficiency, and ease of quantifying uncertainty without relying on additional modelling assumptions.

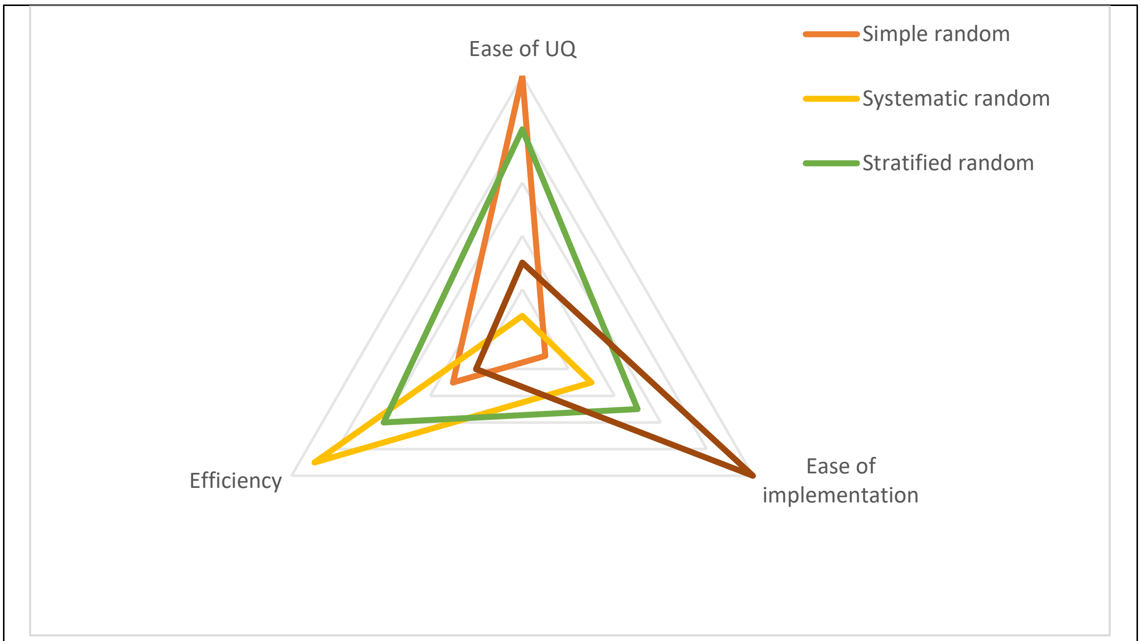


Figure 2.5 *An overview of how different designs compare in the context of estimating accuracy and total prevalence land cover mapping applications.*

For a further discussion on the topic of design choice in land cover mapping applications see [157].

2.2.2 Methods for creating efficient sample designs

As discussed in Section 2.2.1, some sample designs can be more efficient than others depending on the context. This section considers more proactive approaches for creating efficient designs. Here, the literature surrounding these approaches will be grouped under the following headings:

1. Methods for optimising sample designs,
2. The role of aleatoric and epistemic uncertainty in efficient sample design.
3. Accounting for uncertainty in design analysis.

Methods for optimising sample designs

Having decided on a particular type of sampling design, there are still many (sometimes subjective) choices that need to be made. One simple example involves controlling the sample sizes within each stratum under stratified random sampling. As an example, suppose one wishes to estimate the overall accuracy (O_A) of the classifier using simple stratified random sampling. In this case, the maximum likelihood estimate for the overall accuracy is

$$\hat{O}_A = \sum_{i=1}^{m} W_i \, \hat{O}_i = \sum_{i=1}^{m} W_i \frac{k_i}{n_i},$$

where m is the number of strata; W_i is the relative size of stratum i with $\sum_{i=1}^{m} W_i = 1$; n_i is the sample size for stratum i ; k_i is the number of members that were correctly classified. The precision of the estimate can be measured with the variance of the estimate $\hat{O}_A$. In this example, this variance for $\hat{O}_A$ is given by

$$V(\hat{O}_A\,) = \sum_{i=1}^{m} \frac{W_i^2 O_i(1 - O_i)}{n_i}, \tag{2.1}$$

where O_i is the overall accuracy within stratum i. From (2.1) one can see that precision of the estimate, $\hat{O}_A$, is influenced by both the within-strata accuracies, O_i and the relative sizes of each stratum, W_i. Broadly speaking, the optimal choice for n_i under the restriction $\sum_{i=1}^{m} n_i = N$, favours larger stratum where the within stratum accuracies are closer to 0.5 (this is because 0.5 maximises the expression $O_i(1 - O_i)$. Conversely, this means less sampling resources should go to strata that are small or those with either very high (or very low) accuracies. For a more precise balancing of this relationship, one can use methods from non-linear integer programming to optimise size allocation [158]. Such methods can also be extended to include (i) balancing uncertainty reduction across multiple estimates (ii) scenarios where the cost of sampling can vary across the strata, and (iii) restrictions to the minimum and maximum values of each n_i [159], [160], [161].

Whilst each n_i is strictly speaking an integer in this example, dropping this assumption can be computationally convenient. For example, one can make use of Lagrange multipliers to approximate efficient sample designs with closed-form non-integer solutions [162]. This more heuristic approach has been successfully used for applications involving estimating the billing accuracy in insurance claims [163] and estimating total areas in land cover mappings [164].

Similar practices may also be applied to optimise sample designs across more sophisticated design-based methods such as model-assisted estimation [165], [166]. Generally speaking, these methods share the same overall principle as the earlier example with stratified sampling, whereby members with a greater degree of variation will tend to have higher inclusion probabilities when all other factors are fixed. Furthermore, it is possible to give a closed form for (asymptotically) optimal solutions for model-assisted estimation, providing that the model can be correctly specified [59], [167].

When trying to optimise sample designs under model-based approaches, there is a well-established literature on how one may target sampling to manage uncertainty efficiently going back to the late 1950s [168]. Since then, these methods have been extended to include closed-form solutions for generalised linear models [169] and numerical methods for more complex model structures [170], [171], [172], [173].

The role of aleatoric and epistemic uncertainty in sample design

Distinguishing between aleatoric and epistemic uncertainty plays an important role in sample design, as this offers a way of informing users when they are approaching the point of diminishing returns for further sampling. Whilst the idea of distinguishing between aleatoric and epistemic uncertainty (and quantifying their components) is not necessarily new [174], there has been a growing interest in using components of uncertainty in machine learning applications [175], [176]. Recent work has focused on how epistemic and aleatoric uncertainty may be quantified for popular MLTs including Random Forest [177] and neural networks [178], [179].

Techniques from more traditional concepts from statistical inference may also be reframed as discussions related to aleatoric and epistemic uncertainty. As an example, showing the consistency of an estimator [180] is equivalent to showing there is no aleatoric component (or it is infinitely small). Discussions related to power analysis [181] can be reframed as quantifying the rate at which epistemic uncertainty declines as sample sizes increase under a particular type of design.

Separating aleatoric and epistemic components offers a useful way of guiding efficient sample designs that do not rely on optimisation methods. This could be a key advantage under MLTs where it is often difficult to apply traditional optimisation techniques due to the complexity and black-box nature of many MLTs. The potential drawback though is that analysing aleatoric and epistemic components does not give the same explicit suggestions in the same way optimisation methods can.

Accounting for uncertainty in design analysis

When creating sample designs, the efficiency is often dependent on unknown parameter values. For example, in the case of stratified sampling in (2.1), one would need to know each O_i. One expectation of this is in the case of simple linear regression [182] where the optimal designs are based purely on the independent variables, though this is very much an exception that proves the rule. Measures such as aleatoric and epistemic components may also be dependent on unknown parameter values. This can be exemplified in (1.1), where the true value of σ determines the aleatoric component of uncertainty.

When dealing with parameter uncertainty, Bayesian inference is a useful tool for seamlessly folding uncertainty in parameter values into design analysis [183], [184]. Furthermore, stochastic variation and Bayesian inference complement each other well, as by treating unknown parameter values as random variables, the problem of optimising sample designs based on prior information can be treated as a stochastic optimisation problem [185], [186].

In addition to the uncertainty in parameters, stochastic variation from probabilistic sampling and noise components in modelling can also add a degree of uncertainty when formulating how sample design will affect uncertainty. Here, there is a good deal of work related to stochastic optimisation methods that may prove useful in this thesis [187], [188], [189].

Overall, any analysis related to the true efficiency of a sample design will itself be subject to some uncertainty. If this uncertainty is not appropriately accounted for, one runs the risk of placing false confidence in sample designs. In the context of adaptive forms of sampling, one may be especially prone to such a pitfall, as, in early iterations, estimates are likely to be less precise because of the lower sample sizes.

2.2.4 Reflections.

When creating sample designs that can efficiently manage uncertainty, there is much in the current literature to draw upon. From Section 2.2.1, one can see that there are a wide variety of standard sample designs one can potentially choose from. The most appropriate choice of sample design often involves balancing several criteria such as how easy (or costly) it is to implement a design, the ease with which uncertainty can be quantified, and the overall efficiency in terms of uncertainty reduction.

From Section 2.2.2, one can see there are many ways to improve the efficiency of designs through targeted sampling. In addition, one can answer questions such as "*how much reference data are enough?"* by considering the aleatoric and epistemic components of uncertainty. Furthermore, methods of stochastic optimisations pair well with Bayesian inference as a way of incorporating uncertainty into these processes.

For this thesis though, there are still two key remaining challenges, representing gaps in the state-of-the-art.

The first challenge is in identifying a generalisable approach for selecting which type of design to adopt. When selecting types of sample design and reference data, the act of balancing different criteria is often sensitive to context-specific details (see Example 2.5). As one moves on to more advanced scenarios (e.g. using propensity scoring to allow data collected from non-probability sampling in UQ), managing these different considerations can quickly require a substantial degree of domain knowledge. To further complicate matters, generating efficient sample designs requires already knowing (or having a good idea of) the very parameter values that one is trying to estimate.

Heavy reliance on domain knowledge is not ideal in any situation. In the context of machine learning classifiers though, this is an especially substantial issue, as this would take away from one of the main advantages of MLTs, which is their ability to be effective in the absence of domain knowledge. Hence, it would be greatly beneficial to have generalisable methods of choosing the most appropriate sample designs and types of reference data (e.g. useful diagnostic tests) so that one can keep to the general theme of not requiring large amounts of domain knowledge.

The second challenge lies in dealing with the fact that questions over efficient sample designs often involve an exploration of different trade-offs, rather than optimising specific designs. In Section 2.2.2, there were a variety of methods that gave users ways of creating efficient sample designs under specific conditions (e.g. model structures, cost restrictions, types of designs etc.). Whilst these methods can be useful, they are only one part of any solution. Other factors such as ontological uncertainty and sampling convenience are not easily quantifiable and are prone to subjectivity, meaning that there is not always an objectively optimal choice of design. In addition, the choice of model can have a large impact on the relationship between sample design and uncertainty reduction.

Overall, one key lesson from this section is that there is more to managing uncertainty efficiently than optimising sample designs. Often, managing uncertainty efficiently will mean trying to find sample designs that give the best trade-offs between different forms of uncertainty and other subjective elements given a limited number of resources.

2.3 Machine learning techniques in classification problems

The performance of machine learning techniques in classification plays an important role when it comes to quantifying and managing uncertainty, as an accurate classifier can lay a strong foundation for reducing uncertainty efficiently. As an example, one can look back to the stratified random sampling case discussed in Section 2.2.2. Here, the variance for the estimate of the total area, $\hat{O}_A$, in (2.1) will reduce as each O_i approaches 1, which is equivalent to the classifier becoming more accurate. This example is illustrative of a wider theme whereby improving the classifiers themselves can be one way of improving the overall efficiency in uncertainty reduction (see Figure 2.6).

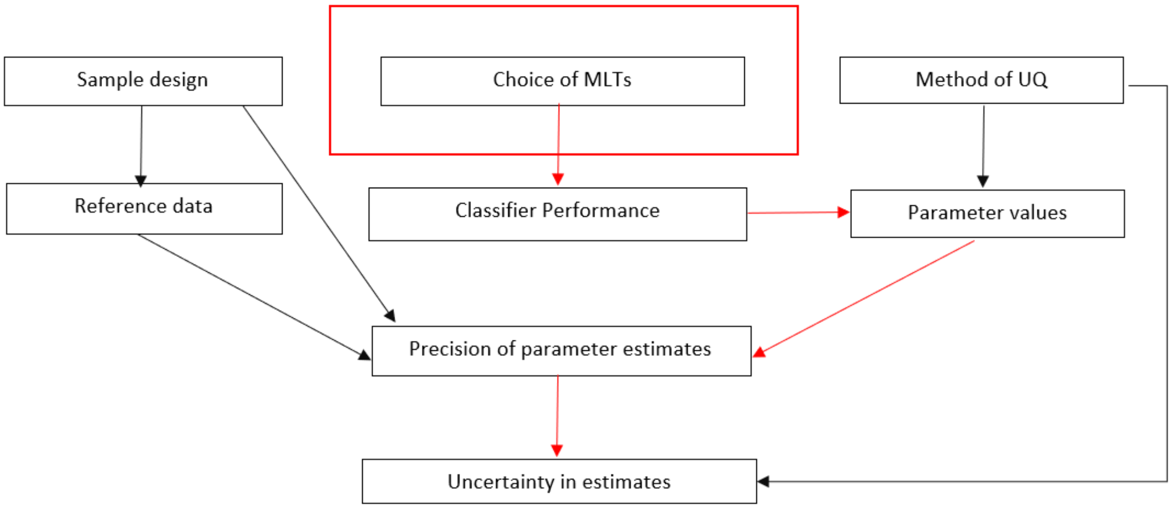


Figure 2.6. *An overview of how the choice of machine learning techniques (red) can impact uncertainty estimates and hence impact the efficiency of uncertainty reduction.*

Overall, there has been a substantial amount of work related to developing and improving machine learning classifiers over the last 70 years. Furthermore, there has been substantial work on how one may do so when under sampling limitations. For this section, it is useful to break this work into the following themes: supervised learning methods (2.3.1), unsupervised, semi-supervised and weakly supervised learning (2.3.2), transfer learning and synthetic data (2.3.3), and methods for selecting MLTs (2.3.4).

2.3.1 Supervised learning

Supervised learning refers to the task of learning a function that maps an input to an output based on example input-output pairs (which will be referred to as fully-labelled data) [190]. In many ways, supervised learning forms the foundation for many machine learning classifiers. Here, one can split supervised learning across three themes: core algorithms, non-parametric and semi-parametric regression, and ensemble learning. The first theme is the set of core algorithms which refers to the set of methods that are commonly cited in machine learning literature and can be viewed as classic MLTs.

The second theme of MLTs is those that are based on non-parametric and semi-parametric regression. These methods differ from the core set in the fact that the approach is closer to traditional statistical modelling.

The final theme of supervised learning discussed in this section is ensemble learning [191]. With ensemble learning, classifiers use a combination of multiple classification algorithms (often from the core set introduced earlier) and are a popular choice for their seeming ability to avoid overfitting and that they are generally easier to implement multiple simple classifiers over a single more complex one [192].

Table 2.7 provides some examples of supervised learning methods from each theme.

Table 2.7 *Examples of supervised learning methods separated across the themes: core methods, non-parametric and semi-parametric regression, and ensemble learning.*

Supervised learning methods	
Core algorithms	
Name	**Source(s)**
Bayesian network classifiers	[193]
Naive Bayes classifiers	[194]
K-nearest neighbours	[195]
Symbolic machine learning	[196], [197]
Logistic regression	[198],
Artificial neural networks	[199], [200]
Support vector machines (SVMs)	[201]
Non-parametric and semi-parametric regression	
Gaussian process models	[202]
Generalised additive models	[203]
Multivariate adaptive regression splines (MARS)	[204]
Kernel regression	[205]
Ensemble learning	
AdaBoost	[206]
Gradient boosting	[207]
Bootstrap aggregation (bagging)	[208]
Subspace partitioning	[209]
Error-correcting codes	[210]
Random feature selections	[211]

Note that these themes are designed to be useful descriptors of supervised learning methods and there is a degree of subjectivity and overlap. For example, the Random Forest classifier [212] is a well-known ensemble method based on decision trees that could arguably be considered a core method these days. Similarly, there is no definitive standard for when a type of non-parametric regression model gains enough popularity in machine learning to cross over into being a core method of supervised learning.

The main motivation for highlighting these themes is that (i) non-parametric and semi-parametric regression models originate from formal statistical modelling, meaning they interact with UQ differently when compared to many core methods and (ii) ensemble methods are often additions to classification techniques which can be separated from the underlying classifiers.

2.3.2 Unsupervised, semi-supervised, and weakly-supervised learning

Unsupervised, semi-supervised, and weakly-supervised learning are all forms of machine learning that aim to improve classifiers by making use of data that is outside the standard pairing of input features and labelled data used to train supervised learning methods.

With unsupervised learning, training data are assumed to be unlabelled; meaning the data contain inputs and no labels. Under semi-supervised learning, training makes use of both unlabelled and labelled data. Weakly supervised learning makes use of weakly-labelled data; meaning there is some sort of labelled data, but it is somehow not of the same quality of manual categorisation that is assumed in supervised learning (e.g. labelling may be imprecise or inaccurate). Figure 2.8 depicts how different forms of supervision in machine learning make use of different types of labelling.

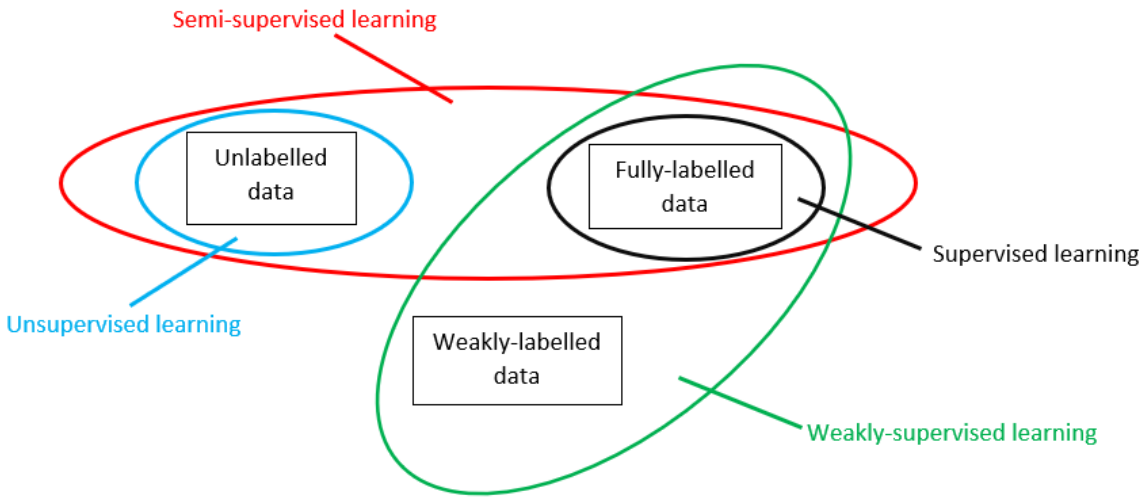


Figure 2.8 *A representation of how forms of supervision in machine learning make use of different types of labelling in reference data.*

From the perspective of efficiently managing uncertainty in machine learning classifiers, it is useful to break the literature surrounding unsupervised, semi-supervised, and weakly-supervised learning across three broad categories: adjusting supervised learning methods through modelling, unsupervised pre-processing, and label reinforcement.

Adjusting supervised learning methods through modelling

Many supervised learning methods can be adjusted to include unlabelled and weakly-labelled data. Many supervised learning methods can combine unlabelled data with fully-labelled data by extending existing objective functions [213]. Some specific examples of extending objective functions to include unlabelled data include support vector machines [214], neural networks [215], Random Forests [216], and Gaussian process models [217].

When dealing with weakly-labelled data, inaccuracy or imprecision may be accounted for by including noise structures in modelling [218]. This approach tends to synergise well with Bayesian inference [219] and MLTs based on non-parametric and semi-parametric regression [220].

Unsupervised pre-processing

Unsupervised pre-processing uses unsupervised learning methods as a means of cleaning or preparing the training data before fitting said data to a separate supervised learning classifier.

One popular class of unsupervised pre-processing are *clustering* methods. Clustering methods aim to automatically detect and associate members of a population with similar input features. Some popular clustering methods in machine learning include Hierarchical clustering [221], K-means [222], and DBSCAN [223]. In classification contexts, the core premise of clustering techniques is that features with similar inputs are likely to belong to the same (or similar) classes. Clustering methods tend to pair well with subspace partitioning methods and can set up the modelling assumptions for semi-supervised classification [224], [225], [226], [227].

Another popular class of unsupervised pre-processing are *dimensionality reduction* methods. With dimensionality reduction methods, the aim is to project the data within the input feature space onto a lower-dimensional manifold that contains as much relevant information as possible to mitigate against the *curse of dimensionality* [228] - a phrase that describes how machine learning becomes exponentially more difficult as the number of input features (or the dimensionality) increases. Popular dimensionality reduction techniques include principal component analysis (PCA) [229], self-organising maps [230], and using neural networks in autoencoding [231].

Label reinforcement

Label reinforcement aims to improve classification by enhancing how unlabelled and weakly-labelled data are used without altering the classification method.

One popular label reinforcement is a multiple-voter approach where one compensates for weak labelling by taking multiple readings of the same instance [232], [233], [234]. Other label reinforcement methods aim to improve weakly-labelled data by identifying points that are likely to be mislabelled so that they can be amended or removed [235].

Another approach to label reinforcement is multi-instance learning [236]. With multi-instance learning, training data are arranged into sets, called bags, and a label is provided for the entire bag. In recent years, there has been a growing area of research on how multi-instance learning can help reinforce unlabelled and weakly-labelled data when there is a limited set of fully-labelled data available [237], [238], [239].

2.3.3 Transfer learning and generating synthetic data

When reference data within a domain is limited due to sampling restrictions, it can be difficult to get enough data to sufficiently train machine-learning classifiers. This section notes two types of approaches that may improve classifiers when training data are limited: transfer learning and generating synthetic data.

Transfer learning

Transfer learning aims to improve machine learning classifiers under limited training data by transferring information from a similar or related problem to a current one. The premise here is that one can train a classifier with fewer reference data by making use of the lessons learned from similar problems (e.g. estimates for model parameters, optional choices for machine-learning structures and hyperparameters etc).

Over the last decade, there has been a noticeable interest in the applications of transfer learning [240], [241], [242], [243] and it has seen use in natural language processing [244], [245], image classification [246], [247] and time series classification.

Synthetic data generation

An alternative way to overcome the problem of training classifiers under limited reference data is to use synthetic reference data to assist training procedures. The core idea of this approach is that synthetic data are much easier to generate than real-world reference data and can be used alongside a relatively small set of real-world reference data. Some popular methods that use synthetic data to improve classifiers include: up-sampling [248], the Synthetic Minority Oversampling Technique (SMOTE) [249], adversarial example generation [250], [251] and generating synthetic reference data based on model simulations [252], [253], [254], [255].

2.3.4 Methods for choosing suitable MLTs

From sections 2.3.1-2.3.3, one can see that there is a diverse set of tools available in machine-learning classification. As this set of tools grows, choosing the most suitable methods in a given situation can seem increasingly more difficult. Over the years, there has been substantial work on choosing suitable machine-learning methods which can be split across the themes of *internal performance assessments*, *surveying*, and *meta-learning.*

Internal performance assessments compare various performance metrics when in the training stage of machine learning. Examples of internal performance metrics include cross-validation [256], out-of-the-bag error rates [257] along with more statistically formal methods such as hypothesis testing [258], [259] and Bayesian inference [260].

Surveying is an approach that evaluates MLTs across different problems to investigate if there are any wider trends regarding the suitability of techniques. Typically, surveying is done at a domain-specific level with examples found in the early detection of lung cancer [261], malware detection [262], and land use and land cover mappings [33].

Meta-learning is based on the idea of learning to learn. [263] describes meta-learning as *"the science of systematically observing how different machine learning approaches perform on a wide range of learning tasks, and then learning from this experience, or meta-data, to learn new tasks much faster than otherwise possible."* In recent years, there has been a particular interest in meta-learning to automate decision-making processes in neural networks and deep learning [264], [265].

It is important to stress that internal performance assessments, surveying and meta-learning are complementary forms of analysis that tend to focus on different levels. For example, internal performance assessments typically focus on one problem but allow the option to explore methods to a greater depth (e.g. where misclassifications are likely to occur, false positive and false negative rates etc.). On the other extreme, meta-learning is often forced to abstract away via meta-features and make inferences at a more general level, which can be a useful practice for shortlisting. Surveying typically lies somewhere in between these two extremes, and it is often easier to conduct a qualitative assessment with surveying when compared to meta-learning.

2.3.5 Reflections

From sections 2.3.1-2.3.4, it is evident that the use of machine learning techniques in classification problems is a well-studied topic that spreads across many subjects. Figure 2.9 summarises the subjects and methods discussed in sections 2.3.1-2.3.4. However, due to the vast literature surrounding the topics, it is important to note that sections 2.3.1-2.3.4 and Figure 2.9 are by no means an exhaustive list of available methods in machine learning classification.

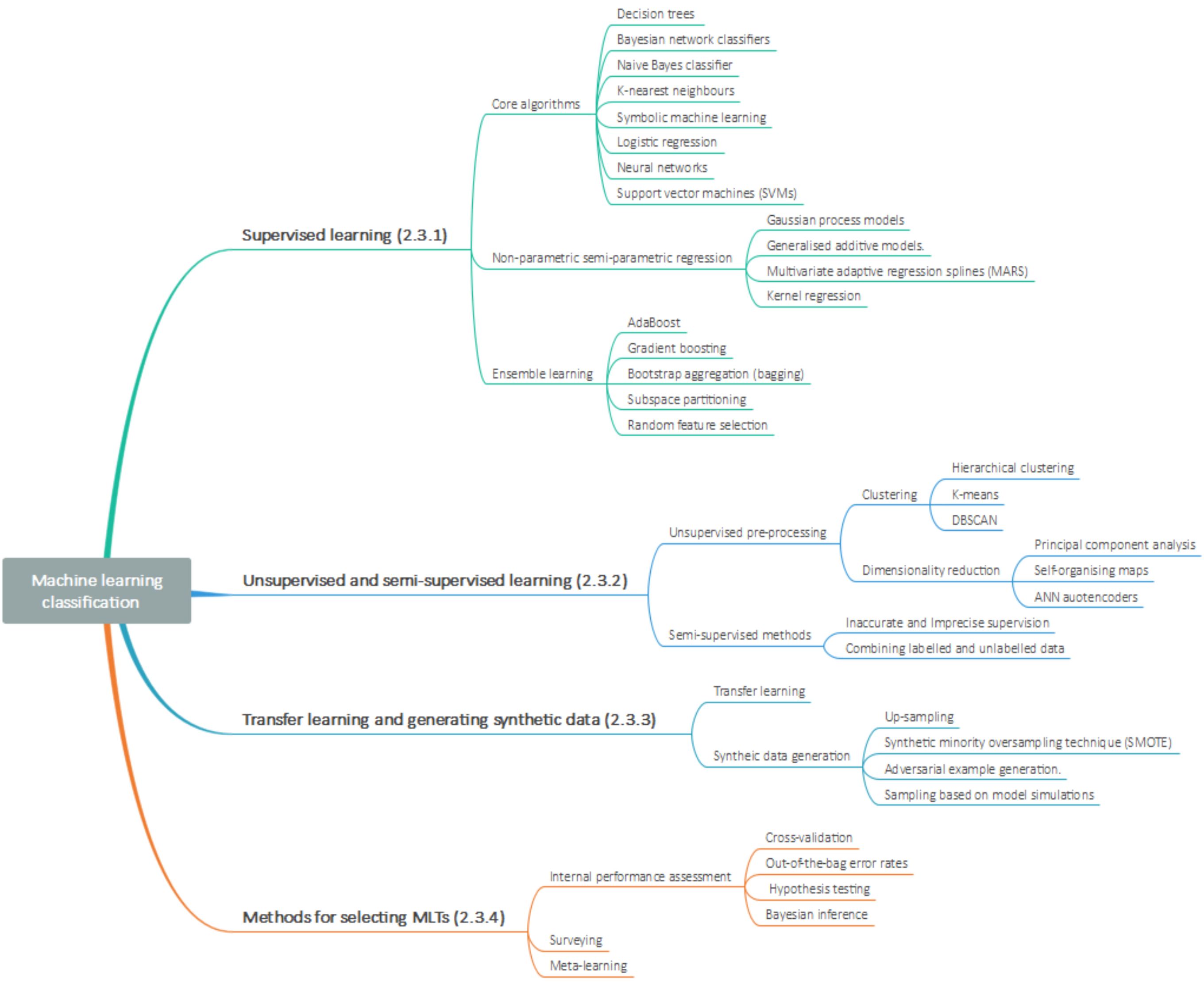

Machine learning classification
Supervised learning (2.3.1)
Core algorithms
Decision trees
Bayesian network classifiers
Naive Bayes classifier
K-nearest neighbours
Symbolic machine learning
Logistic regression
Neural networks
Support vector machines (SVMs)
Non-parametric semi-parametric regression
Gaussian process models
Generalised additive models.
Multivariate adaptive regression splines (MARS)
Kernel regression
Ensemble learning
AdaBoost
Gradient boosting
Bootstrap aggregation (bagging)
Subspace partitioning
Random feature selection
Unsupervised and semi-supervised learning (2.3.2)
Unsupervised pre-processing
Clustering
Hierarchical clustering
K-means
DBSCAN
Dimensionality reduction
Principal component analysis
Self-organising maps
ANN auotencoders
Semi-supervised methods
Inaccurate and Imprecise supervision
Combining labelled and unlabelled data
Transfer learning and generating synthetic data (2.3.3)
Transfer learning
Syntheic data generation
Up-sampling
Synthetic minority oversampling technique (SMOTE)
Adversarial example generation.
Sampling based on model simulations
Methods for selecting MLTs (2.3.4)
Internal performance assessment
Cross-validation
Out-of-the-bag error rates
Hypothesis testing
Bayesian inference
Surveying
Meta-learning

Figure 2.9. *A summary of the topics and methods in machine learning classification discussed in sections 2.3.1-2.3.4*

From the perspective of managing uncertainty efficiently, the choices of MLTs in classification play an important role, as the quality of a classifier can impact overall efficiency. Hence anything used to improve or build upon the quality of classifiers can potentially help in reducing uncertainty efficiency in machine learning classification.

The most direct option for improving the classification method may be to improve the supervised learning methods. Examples of this include using different activation functions in neural networks [266], [267] or using different kernel functions in SVMs [268], [269]. However, it is not clear if there is much scope for improving supervised classification methods by simply tweaking hyper-parameters or cost functions at this stage of development.

Complementary methods such as semi-supervised and weakly supervised learning (2.3.2), transfer learning or using synthetic data (2.3.3) may look promising for improving classification methods, especially when ground-truth reference data are limited. With such methods though, there is no guarantee of improving classifiers. Some authors have reported that such methods can hurt performance. Examples of this can be found in semi-supervised algorithms [270], [271], [272], inaccurate supervision [218], [273], [274], multi-instance learning [275], and transfer learning [242], [276], [277], [278]. Hence, to apply such methods consistently, one would ideally want to know when such methods are likely to be beneficial. The idea of using meta-learning techniques (2.3.4) to automate the process of choosing suitable MLTs and tuning hyperparameters has also been explored in recent years [279], [280], [281]. However, the degree that these methods can consistently select the better-performing classifiers is an open-ended question at this stage.

Regardless of whether any of the methods discussed in this section can improve classification algorithms though, there is more to selecting MLTs than raw performance measures such as accuracy and precision. In many circumstances, there are other factors to consider such as the ability to handle missing data [282], [283], sensitivity to hyper-parameter choices [284], [285], reliance on pre-processing techniques [57], [286], or their ability to deal with categorical data etc.

For example, one might be able to show that, under the right kernel structure, SVMs can outperform Random Forests under limited training data in a context where the predictors are fully available and continuous. However, in a context where data may be missing from inputs (e.g. some questions in a survey left unanswered) and one is dealing with discrete inputs, Random Forests are still likely to be more suitable than SVMs.

From this discussion, it becomes clear that, whilst it may theoretically be possible to manage uncertainty more efficiently by improving the quality of the classifiers themselves, the general topic of improving classifiers with MLTs is a mature one. Hence, any improvements in uncertainty reduction via improving machine learning classification may only be incremental.

Furthermore, improving the performance of machine learning classifiers would only serve as one potential component when considering the wider aim of quantifying and managing uncertainty more efficiently. Many MLTs discussed in this section fail to meet the sampling and modelling requirements discussed throughout Section 2.1 and 2.2. For example, transfer learning may be a useful practice for producing more accurate classifiers given a limited amount of training data, but it is not clear how transfer learning would be formally integrated into methods of UQ (without relying on some heavy modelling assumptions).

This thesis proposes that a key gap in the literature lies in developing tools for quantifying and efficiently managing uncertainty that is agnostic to the choice of MLT.

This is not to say that the choice of MLT is not important for reducing uncertainty. One can anticipate that with time, existing MLT algorithms will eventually be replaced with more accurate and precise methods. Rather, this thesis proposes that given the diverse range of methods available, the complexity of the data sets typically involved in machine learning problems, and domain-specific restrictions, any framework for managing uncertainty efficiently will need to have the flexibility to deal with a diverse set of plausible MLTs and associated methods.

To phrase this another way, the gap in the current research does not lie so much in making or selecting better MLTs these days but instead lies in how one quantifies (and efficiently manages) uncertainty for a given machine learning classifier approach and context.

2.4 The Intersection of MLTs, UQ, and Reference Sampling

This section reviews the current literature for work that combines uncertainty quantification and reference sampling with machine learning classification. More specifically, this section focuses on the following topics: reference sampling for MLTs (2.4.1), combining Bayesian inference and MLTs (2.4.2), and interpretability and explainability in MLTs (2.4.3).

2.4.1 Reference sampling for MLTs

MLTs often require large volumes of training data to be effective. Whilst some methods can help mitigate this problem (see Section 2.3 for further details), this subsection focuses on how to collect the data more efficiently to begin with. In other words, Section 2.3 focuses on how one may train MLTs efficiently once given the reference data, whereas this subsection focuses on how one might sample the reference data more efficiently.

One approach to sampling reference data more efficiently is to make use of reference data that has been volunteered by third parties (this is sometimes referred to as citizen science) [287]. Using volunteered data as a cost-effective means of training classifiers has been gaining interest in many domains including archaeological prospection [288], collecting ground-truth data for land cover mappings [289] and earth science applications [290]. Bayas *et al* [291] took this a step further and considered how to combine volunteered data and targeted sampling in land cover mapping applications by incentivising users to target areas through gamification.

For a more controlled approach to creating efficient designs, a subfield of machine learning known as active learning [292], [293], [294], [295] may also be useful. Ren *et al* [295] describe active learning as:

"Active learning (AL) attempts to maximize a model's performance gain while annotating the fewest samples possible."

Over the years, active learning has been successfully applied to many popular MLTs including linear classifiers [296], support vector machines [297], and neural networks via improving stochastic gradient descent methods [298], [299], [300]. Furthermore, there has been recent work combining active learning and Bayesian neural networks to produce deep active Bayesian learning [301], [302].

The real-world benefits of active learning have been seen in material science [303], text labelling applications [304], [305] and object detection methods [306].

Although these approaches help in producing more cost-effective training methods for training MLTs, some challenges remain. Firstly, the use of volunteered data may only be a viable option for specific domains. For example, Scheibein *et al* [307] highlight potential ethical and reliability concerns in volunteered science data in addictions and substance-use research.

Secondly, the challenges surrounding UQ under non-probability sampling discussed in Section 2.2.3 still hold when considering volunteered data and active learning. Under the framing set out in Section 2.2.3, volunteered data may be viewed as a specific case of non-probability sampling. Likewise, active learning techniques may be viewed as heuristic optimisation methods.

2.4.2 Combining Bayesian inference and MLTs

Over the last decade, there has been a noticeable amount of work on combining Bayesian inference with popular MLTs such as support vector machines (SVMs) [308], [309] and neural networks [178], [310]. Theoretically, MLTs such as SVMs and neural networks can be treated as any other parameterised model, meaning that one should be able to apply Bayesian techniques to them.

However, the complexity and high dimensionality involved in many MLTs creates heavy computational demands when generating posterior distributions, making Bayesian inference for MLTs difficult to realise in practice. Recently there has been a growing interest in methods that can help overcome the issues of these computational demands so that one can benefit from the advantages of machine learning and Bayesian inference simultaneously. For SVMs, Wenzel *et al*. [311] took inspiration from previous work in Gaussian process models [312], [313] to use stochastic variational inference and inducing points when generating posterior distributions.

For neural networks, some examples include using bootstrapping methods [314], truncating the problem to focus only on output layers [315], and using gradient descent methods to approximate posterior distributions [316], [317].

Even though reducing computational demands for posterior generation is an important step towards benefiting from the advantages of both Bayesian inference and MLTs, these methods alone will not be enough to address the wider problem of efficiently managing uncertainty in MLTs.

This is because any method that reduces the computational demands in posterior generation can only ever make the calculations in UQ easier, but they do not address whether these quantifications of uncertainty are precise enough or even valid. To address the question of validity in UQ under Bayesian neural networks, one must still carefully consider the sample design for reference data used in training alongside any other modelling assumptions such as prior distributions and hyperparameters set by the user.

2.4.3 Interpretability and Explainability in MLTs

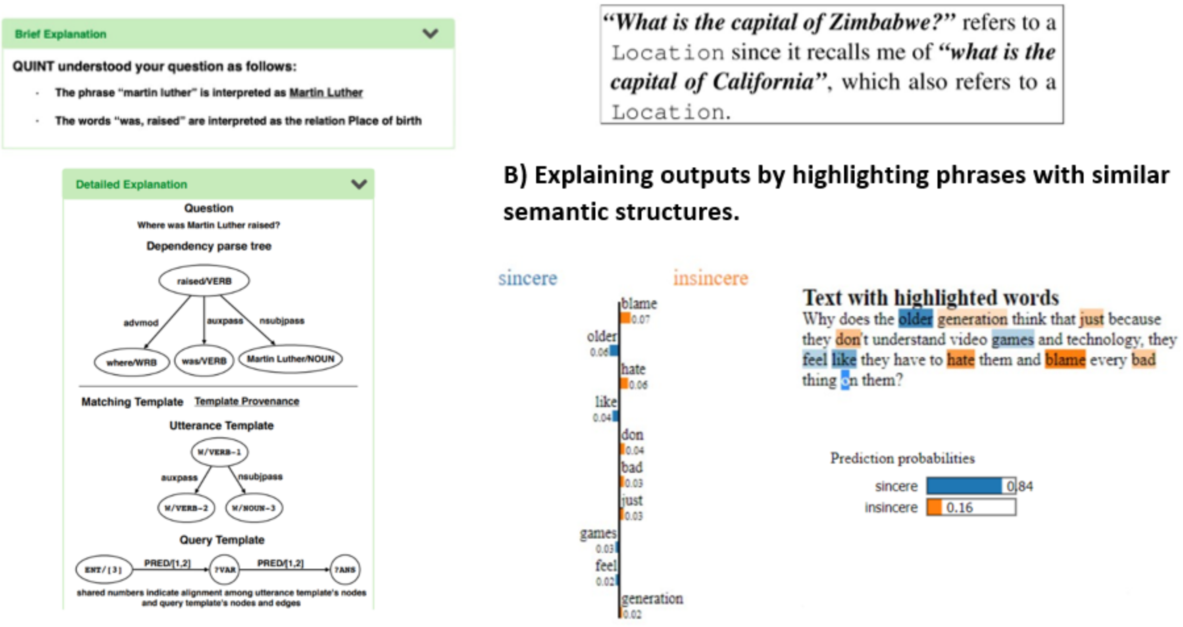
One of the major challenges for UQ in machine learning stems from the problem that many MLTs are black-box in nature, which causes issues with ontological uncertainty as verifying modelling assumptions in classifiers becomes difficult when it is unclear how these techniques are operating in the first place.

When trying to address the issues of UQ in black-box MLTs, it is useful to consider the ideas of interpretability and explainability. A popular definition for interpretability presented by Miller defines interpretability as *“the degree to which a human can understand the cause of a decision”* [318]. Explainability refers to the degree that the internal logic and mechanics inside a machine-learning system are understood [319].

Although related (and often used interchangeably), there is a subtle difference between interpretability and explainability that is important for the discussion of UQ in MLTs. Explainability is necessary for UQ because one needs some understanding of the internal mechanics of a machine learning classifier to know what modelling assumptions are being made. Interpretability is necessary for handling ontological uncertainty in UQ as there needs to be a degree of human-level understanding within a classifier if humans are to decide if any modelling assumptions are appropriate.

As machine learning has become more popular, there has been a rising interest in methods that make MLTs more interpretable and explainable [319], [320]. Many approaches focus on adding interpretability to existing MLTs. Some popular methods here include using importance measures and sensitivity analysis to indicate which features are influencing machine learning outputs the most [321], [322], [323], [324], [325], [326], creating adversarial examples to assess the robustness of classifications [327], [328], [329], and emulating MLTs with easier to understand models such as Decision Trees [330], [331] or localised linear models [332], [333], [334].

Approaches for making MLTs more interpretable have become popular in natural language processing [335], [336] and image classification problems [337], [338], [339], [340]. Figure 2.10 provides examples of methods that aim to add interpretability to MLTs in NLP problems and Figure 2.11 provides examples for image classification applications.



Brief Explanation

QUINT understood your question as follows:

- The phrase "martin luther" is interpreted as Martin Luther
- The words "was, raised" are interpreted as the relation Place of birth

Detailed Explanation

"What is the capital of Zimbabwe?" refers to a `Location` since it recalls me of ***"what is the capital of California"***, which also refers to a `Location`.

B) Explaining outputs by highlighting phrases with similar semantic structures.

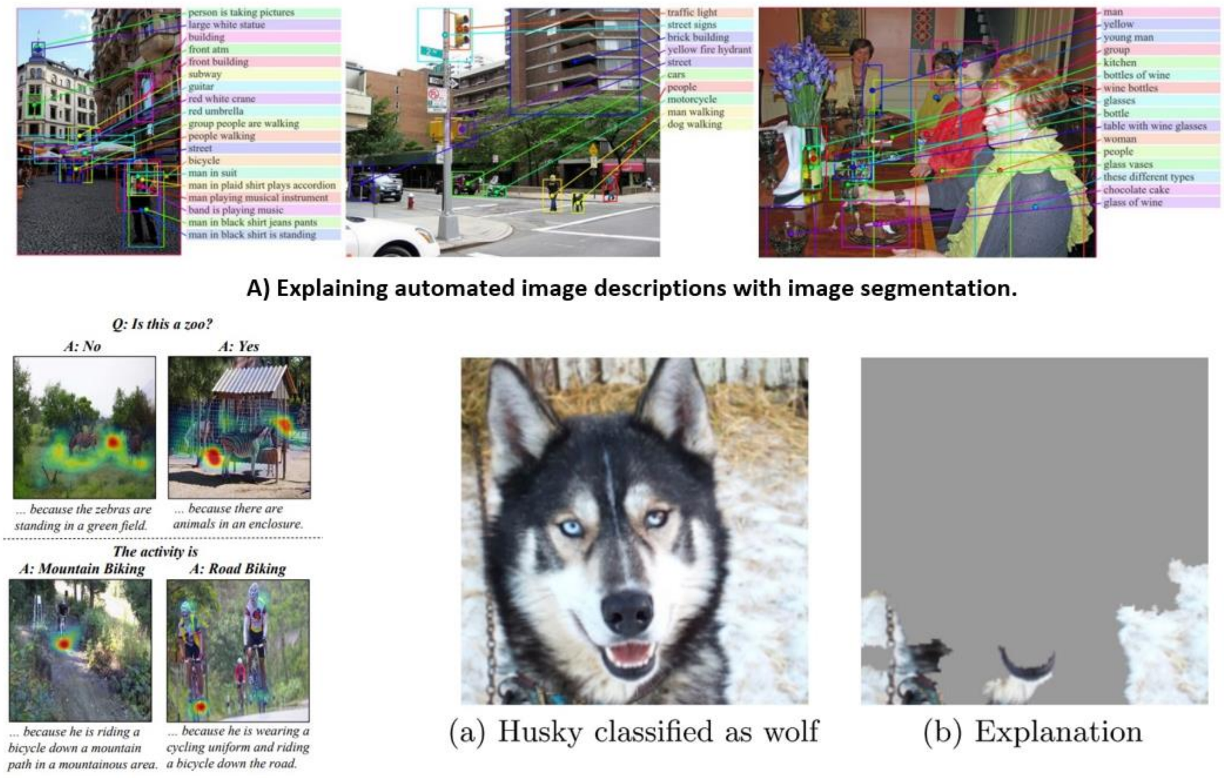
Text with highlighted words

Why does the older generation think that just because they don't understand video games and technology, they feel like they have to hate them and blame every bad thing on them?

A) Non-technical and detailed explanations provided by QUINT.

C) LIME example when classifying if a statement is sincere.

Figure 2.10. *Examples of approaches that aim to add interpretability to MLTs in NLP applications. A) Non-technical and detailed explanations provided by QUINT for the automated response to the question "Where was Martin Luther raised?"* [341] *B) Explaining outputs by highlighting similar phrases with kernel-based methods* [342] *C) An example of LIME being used to explain why the classifier believes that the statement is sincere* [319].



A) Explaining automated image descriptions with image segmentation.

B) Highlighting influential areas in images to explain classifications.

C) Highlighting influential areas in images to detect weaknesses in classifiers.

Figure 2.11. *Examples of approaches that aim to add interpretability to MLTs in image classification applications. A) Using image segmentation to explain MLTs that generate automatic descriptions for images* [343] *B) Using ACT-X to highlight influential areas in images and provide text-based explanations* [342] *C) An example of highlighting influential areas to detect weaknesses in classifiers (in this case falsely classifying a husky as a wolf based on the presence of snow alone)* [319]

Whilst approaches such as those in Figures 2.10 and 2.11 may be useful when developing machine learning classifiers and building user trust [344], [345], there are major limitations when it comes to quantifying uncertainty. Rudin [346] discusses the limitations of using post-hoc approaches to add interpretability and explainability at a general level where two main issues are particularly relevant to uncertainty quantification.

The first issue is these methods often fail to provide enough explainability that is needed for UQ. For example, Figure 2.11 illustrates how one can highlight influential areas of an image as part of explanations. However, there is no immediate way of going from this type of information into a formal statement that quantifies uncertainty as it is still unclear how the machine learning classifiers are operating.

The second issue is determining how faithful any emulations are to the original classifier. Ultimately, any emulation of a machine learning classifier is a model. As such, uncertainty in emulations themselves will need to be accounted for when quantifying uncertainty. If not careful, one can easily fall into the trap of simply shifting many of the black-box elements of the original classifier to the model that links the classifier and emulated output without addressing the core issue which is to quantify uncertainty in the original machine learning classifier.

Some authors have experimented with MLTs that are built with interpretability and explainability in mind from the beginning. These methods are often referred to as intrinsic methods. Some examples of intrinsic methods include: supersparse Linear Integer Models (SLIM) [347] that are restricted to simple functions such as additions, subtractions, and multiplications of input features to keep predictions more interpretable; GA^{2}M [348] which is based on generalised additive models (GAMs) with pairwise interactions; and generalised linear rule models [349] which use generalised linear models to create rule-based classifiers.

Although intrinsic methods like these may help in overcoming issues of interpretability and explainability in MLTs, there is currently a debate over whether this comes at the expense of performance. Rudin [60] argues that it is possible to use intrinsic methods without suffering a substantial penalty to performance in high-stake decision-making, whereas other authors point out that there has been a lack of interest in intrinsic machine learning applications such as NLP and computer vision in recent years as they tend to lack the performance of current black-box methods [29].

As discussed in Section 2.3, the performance of the classifier plays a pivotal role in how much (and how efficiently) uncertainty can be reduced with reference sampling. Hence, the extent of any performance drop in intrinsic methods must be carefully considered.

2.4.4 Reflections.

As machine learning classifiers become more popular, it is only natural that one would want to take methods of quantifying and managing uncertainty through reference sampling (discussed throughout sections 2.1 and 2.2) and combine them with MLTs (discussed in Section 2.3)

From the current literature, this section identified three areas of work that are useful when managing uncertainty in MLTs efficiently. These areas are:

(i) Reducing the training cost of MLTs by using cheaper forms of reference sampling (e.g. using volunteered data) or targeting designs to focus on areas that train MLTs more efficiently (i.e. active learning).

(ii) Quantifying uncertainty in MLTs by treating them as parametrised models and using Bayesian inference.

(iii) Improving the interpretability and explainability of MLTs to address issues related to ontological uncertainty in MLTs.

However, integrating MLTs into the dynamic between uncertainty management and reference sampling is not without its challenges. Often, integrating MLTs results in work where one can either quantify uncertainty in MLTs or use different forms of reference sampling to train MLTs efficiently, but not both. The main reflection from this section is that if one is to develop a framework for quantifying and efficiently managing uncertainty in MLTs, one will need to consider the topics of MLTs, reference sampling, and uncertainty quantification simultaneously, yet a lot of the current literature focuses on at most two of these three topics at a time. Hence, there is a substantial opportunity for this thesis to focus on the intersection of machine learning, UQ, and reference sampling.

2.5 Summary

When managing the trade-offs between uncertainty and the costs of collecting reference data for classifiers built using MLTs, three fundamental questions need to be considered.

1. How should uncertainty be quantified?
2. How should reference data be sampled?
3. How does the introduction of MLT classifiers affect any answers to 1 and 2?

Overall, there is a lot in the current literature that focuses on questions 1 and 2, and a growing focus on answering question 3 as interest in MLTs has increased over the years.

Focusing on question 1 first, a review of the literature indicates that there are many plausible ways to quantify uncertainty, each with different advantages and drawbacks. One key choice a user will need to make is whether to base UQ on a frequentist or Bayesian perspective. Another key choice is between the design and modelling dependencies (see Section 2.1.1 for further details). When discussing the suitability of different UQ methods, one will need to ask questions such as "*To what extent do I believe the modelling assumptions used in this method of UQ hold?*" and "*Do I believe the data here were (or can be) collected in a way that meets its design requirements?*". Generally speaking, Bayesian inference has been gaining popularity in the last two decades and offers a generalisable approach to UQ via Monte Carlo simulation. However, there are still applications where one may prefer frequentist methods (e.g. model-assisted estimation), so the choice here is not an objectively clear one. As for the choice between modelling and design dependencies, there is a lot in the current literature that focuses on the suitability of different dependencies at a domain level. However, it is difficult to extrapolate beyond domain-specific analysis here, as the suitability of different dependences can be sensitive to context-specific factors. For example, obtaining a simple random sample may be easy in one domain, but nearly impossible in another.

In relation to question 2, there is a good body of work related to various types of probability and non-probability designs one can employ when collecting reference data (see Section 2.2.1 for further details). There is also a substantial amount of work-related to how one might make sample designs more efficient via optimisation and analysis of aleatoric and epistemic components of uncertainty. As one becomes more familiar with this literature, it becomes clear that the topics of uncertainty quantification and efficient sample designs are highly interconnected. Some noteworthy examples of this interconnectivity include (i) methods that use propensity scoring will require some modelling dependencies, (ii) Bayesian inference offers a generalisable method for propagating uncertainty in optimisation methods, (iii) design-based methods will not be suitable if some areas are inaccessible (i.e. areas have inclusion probabilities of 0). Note, these examples are not an exhaustive list of such overlaps but rather an illustration of how the method of UQ and choice sample design cannot be viewed in isolation.

Considering question 3, the introduction of MLTs affects the dynamic between uncertainty and reference sampling in three major ways. Firstly, MLTs offer a way of creating accurate classifiers without the need for a large degree of domain expertise. These performance gains will typically reduce noise components, which help in reducing uncertainty efficiently. The second factor is that MLTs tend to require larger sample sizes in training to be effective, which may add a noticeable burden to sampling costs, especially if one is relying solely on probabilistic designs. The third factor is that many MLTs are black-box in nature, as they often lack interpretability or explainability. This can make it difficult to verify or trust methods of UQ that rely on modelling assumptions when using MLT classifiers, potentially limiting the types of UQ one can employ.

There has been a substantial amount of work that aims to tackle or mitigate these factors. For example, there is substantial research on how MLTs may be improved and trained more efficiently along with substantial research on making MLTs more transparent and suitable for UQ. In particular, there has been a lot of focus on the role of transfer learning, active learning, and Bayesian deep learning in recent years. However, much of this work fails to consider how the interconnected nature of UQ and sample design may apply to MLTs, which can create conflicts between many popular methods.

It is from these conflicts that one comes to realise that there is a gap in the literature that focuses specifically on the interaction between UQ and reference sampling in the context of MLTs. Addressing this gap provides an opportunity to enhance MLTs in a way that goes beyond incremental performance gains, but rather addresses the issue of trying to build trust in MLTs through UQ without needing overly expensive sample designs.

Chapter 3 A Framework for Adaptive Sampling

3.1 Introduction

From the review of the literature in Chapter 2, it is clear that the problem of managing uncertainty efficiently under machine learning classifiers will involve a balancing act between uncertainty, practical restrictions in sampling, and the types of machine learning techniques used. However, balancing these choices involves dealing with many interconnected challenges. Some key examples of these challenges include:

- The decision on how to quantify uncertainty.
- Sampling restrictions do not always follow neatly defined objective functions.
- Uncertainty is often tied to the performance of the classifier.
- The fact that machine learning often calls upon a wide range of perspectives and ideas that are not always designed with UQ in mind.
- The breadth, depth, and evolving nature of the machine learning literature.
- Uncertainty in factors that drive these choices (e.g. uncertainty in parameter values, the validity of assumptions in models etc.).

The relationship between uncertainty in machine-learning classifiers and reference sampling generally depends on four factors. These factors are domain-specific features (e.g. restrictions in sample designs, types of data available etc), the choice of machine learning techniques, how uncertainty is quantified, and the sample design.

Figure 3.1 illustrates how these four factors may be viewed as trying to navigate branches in decision trees. Under this view, managing uncertainty efficiently can be seen as trying to find suitable paths through this decision tree (i.e., finding orange paths) given a set of domain-specific features (outlined in blue).

However, finding these efficient paths can be difficult given the wide variety of approaches to UQ, and MLTs available. One option here would be to simplify the process by fixing the MLTs or methods of UQ (outlined in green), but this may narrow the focus too much and does not help a user when these specific methods do not align well with the current problem.

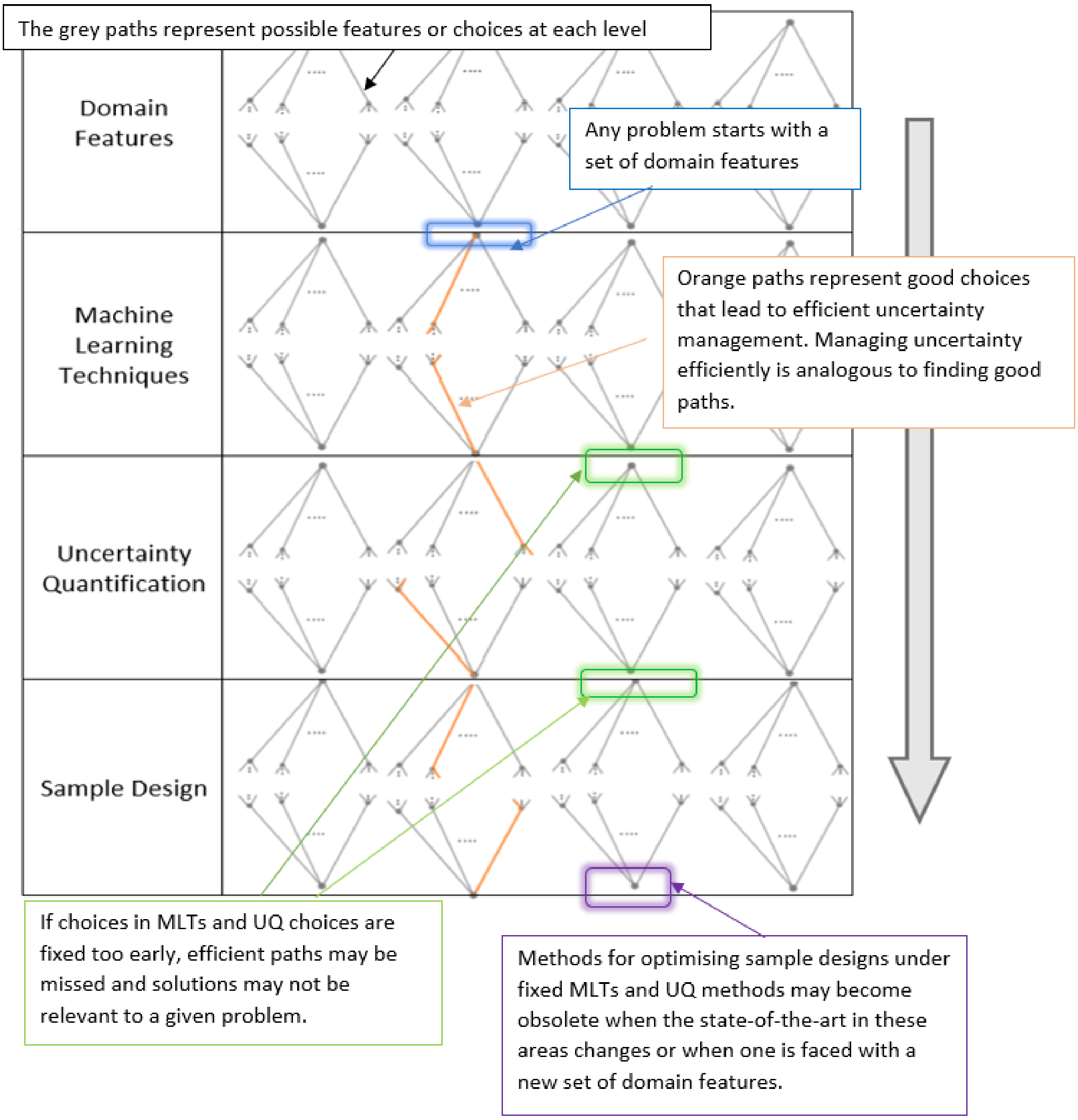

The grey paths represent possible features or choices at each level
Domain Features
Machine Learning Techniques
Uncertainty Quantification
Sample Design
Any problem starts with a set of domain features
Orange paths represent good choices that lead to efficient uncertainty management. Managing uncertainty efficiently is analogous to finding good paths.
If choices in MLTs and UQ choices are fixed too early, efficient paths may be missed and solutions may not be relevant to a given problem.
Methods for optimising sample designs under fixed MLTs and UQ methods may become obsolete when the state-of-the-art in these areas changes or when one is faced with a new set of domain features.

Figure 3.1 *A visual representation of the problem of managing uncertainty efficiently under machine learning classifiers using decision trees. Here, decision trees are made up of domain features and choices related to MLTs, UQ and sample design.*

As the scale and complexity grow across these four factors, the idea of developing a system which can proscribe the best combinations of sample design, UQ and MLT in given problems seems less and less feasible. These problems here are only further compounded by subjective elements in UQ, the ever-changing nature of MLTs, and sensitivity to domain-specific features.

Instead, this thesis proposes a different approach for managing uncertainty efficiently under machine learning classifiers that focuses on an adaptive sampling framework which aims to be agnostic to the choices of UQ and MLTs.

The motivation for this framework is based on two fundamental ideas. The first idea is that adaptive sampling can offer a more consistent means of generating efficient sample designs. Typically, generating efficient sample designs requires knowledge of unknown variables or parameters. This can cause issues when there is an imprecise or inaccurate understanding of these variables, as this can drastically alter how the perceived cost-benefit of a sample design analysis matches reality. With adaptive sampling though, previous iterations are used to update the understanding of these variables and parameters. Consequently, one can build the evidence needed to create more efficient sample designs throughout multiple iterations (and adjust when necessary).

The second idea is that if a framework is agnostic to the choice of MLTs and the method of UQ, it becomes robust and effective in managing uncertainty. The reasoning for this is that an agnostic framework gives users the option to replace MLTs with better ones or change the method of UQ depending on which modelling and design assumptions one is willing to accept. An agnostic framework will also apply to a wider range of domains as it creates less reliance on domain-specific features that affect the viability of different machine learning and UQ choices.

Figure 3.2 illustrates this idea from the decision tree perspective introduced in Figure 3.1. Here, a framework that is agnostic to the choices in MLTs and UQ is represented by wide green bands across these levels.

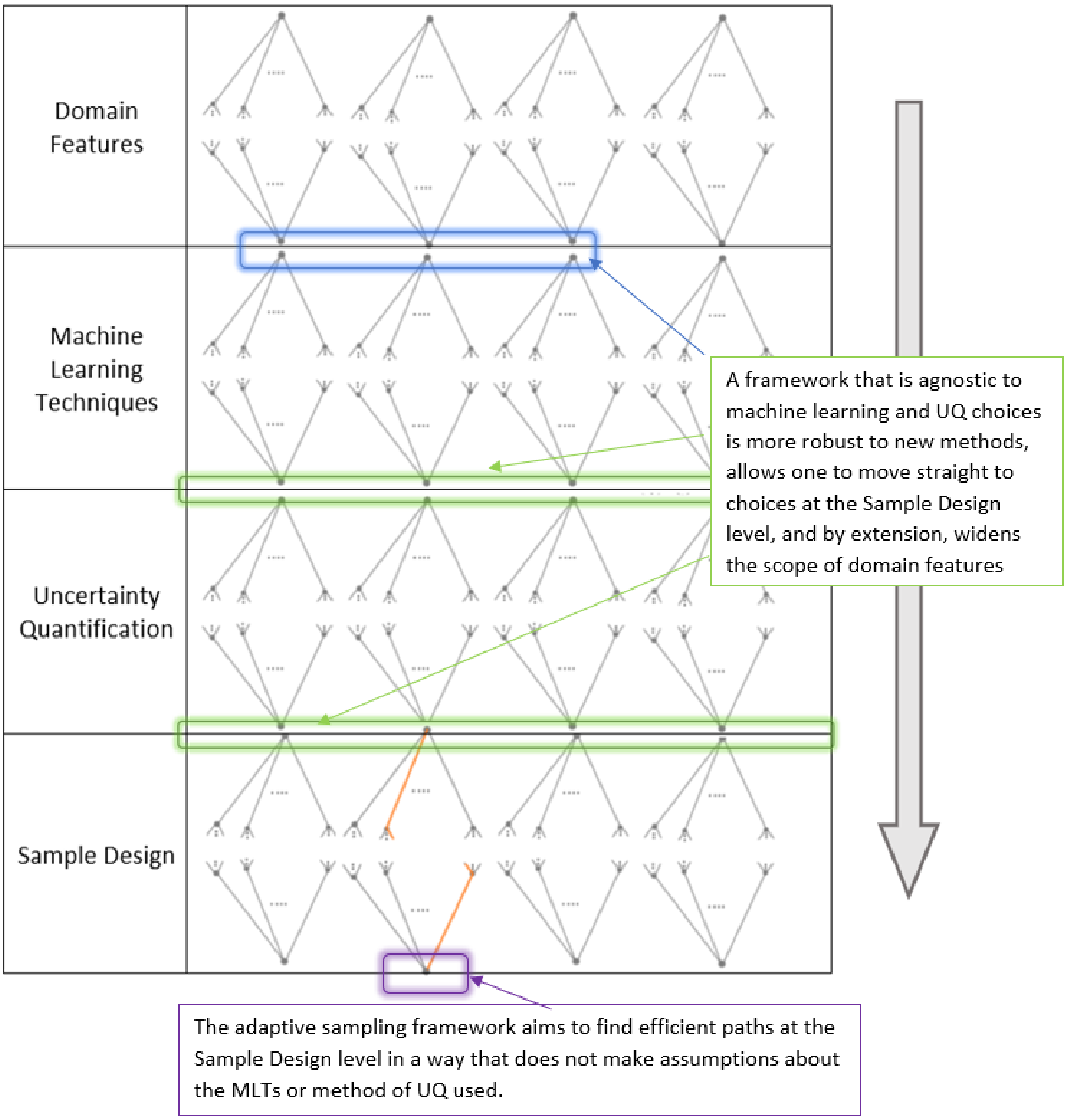

Domain Features
Machine Learning Techniques
Uncertainty Quantification
Sample Design
A framework that is agnostic to machine learning and UQ choices is more robust to new methods, allows one to move straight to choices at the Sample Design level, and by extension, widens the scope of domain features
The adaptive sampling framework aims to find efficient paths at the Sample Design level in a way that does not make assumptions about the MLTs or method of UQ used.

Figure 3.2 *The motivation behind an adaptive sampling framework that is agnostic to machine learning and UQ choices and how it relates to the decision tree perspective introduced in Figure 3.1*

The remainder of this thesis will concentrate on developing an adaptive sampling framework that is designed to help manage uncertainty efficiently in MLTs without being reliant on specific forms of UQ or MLTs. To evaluate this framework, the following criteria will be used:

- **The ability to manage uncertainty efficiently under design restrictions**. This criterion describes the extent to which a framework can help manage uncertainty efficiently under various restrictions or conditions within sample designs. Some of these restrictions may be the total sample size, overall costs, and restricted or unavailable members within a population.

- **Generalisability**. This criterion describes how reliant a framework is on specific forms of UQ or MLTs. A framework will be high in generalisability if there are few dependencies on the types of UQ or MLTs.

The purpose of this chapter is to introduce the core components of the proposed adaptive sampling framework and is structured as follows: Section 3.2 introduces the four phases in the adaptive sampling framework and the methods used to populate the framework. Section 3.3 gives a worked example of the proposed framework. Finally, Section 3.4 summarises the framework and how the practices introduced in 3.3 interact with the four phases of adaptive sampling.

3.2 Establishing the adaptive sampling framework

3.2.1 Overview

To better understand the challenges and opportunities in adaptive sampling, this thesis will introduce a strategic-level overview of adaptive sampling consisting of four stages (see Figure 3.3). These four stages are:

Updating the sample: The act of collecting a new sub-sample based on a specified sampling design and combining it with any previous subsamples.

Updating uncertainty: The act of quantifying the uncertainty for predictions using the total available sample.

Design proposal: The act of generating sample designs for the next sub-sample that are likely to be beneficial (e.g. optimal, cost-effective etc.) based on the currently available sample.

Design assessment: The act of assessing any proposed sample designs based on the current information and deciding upon a sample design for the next sub-sample (note the option of no further sampling is always one proposal here).

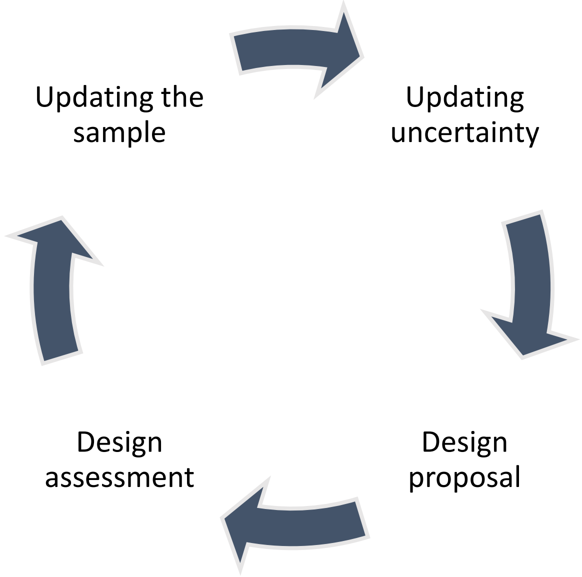


Figure 3.3. *The key stages of adaptive sampling are represented as an iterative process.*

From this strategic-level overview, one can begin to populate the adaptive cycle with tactical-level methods that are designed to pass through these stages without making many assumptions related to the choice of MLT or the method of UQ. The remainder of this section will focus on introducing and motivating such methods. More specifically, this section will introduce and motivate the following practices (i) using Bayesian inference (3.2.2); (ii) using the predictors in a model as a basis for targeted sampling (3.2.3); (iii) quantifying aleatoric and epistemic components of uncertainty (3.2.4); (iv) predicting the likely effects of further sampling (3.2.5). An overview of where these methods interact with the adaptive sampling cycle is provided in Figure 3.4.

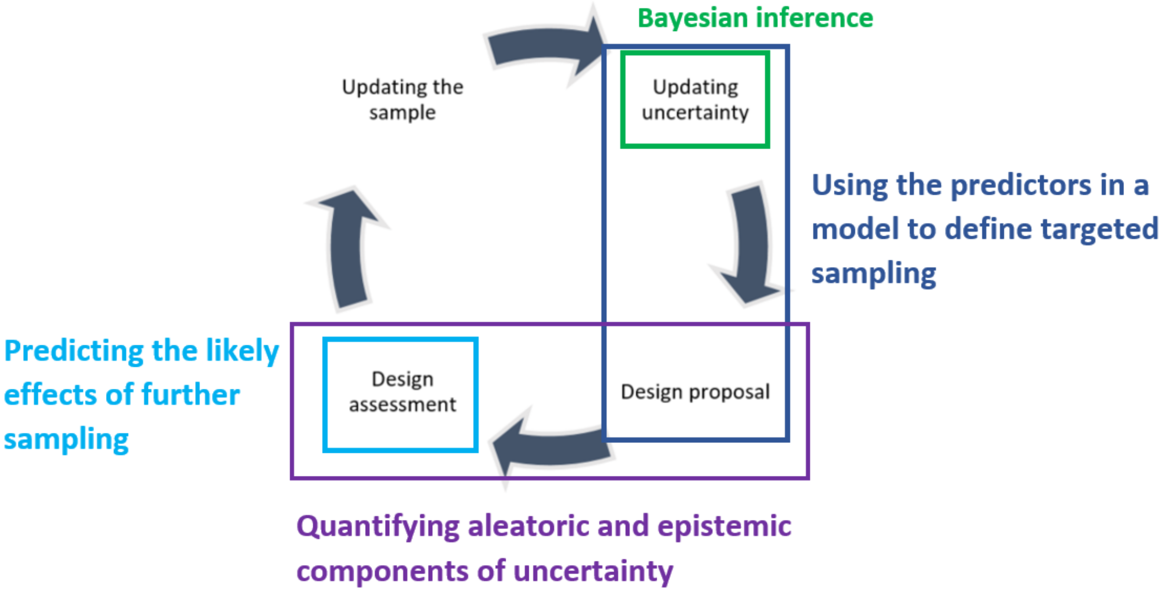

Bayesian inference
Updating the sample
Updating uncertainty
Using the predictors in a model to define targeted sampling
Predicting the likely effects of further sampling
Design assessment
Design proposal
Quantifying aleatoric and epistemic components of uncertainty

Figure 3.4. *A summary of how the methods introduced in Chapter 3 are expected to interact with the adaptive sampling cycle.*

3.2.2 Using Bayesian Inference for UQ in adaptive sampling

Section 2.1.2 introduced two forms of inference for UQ, frequentist and Bayesian. From an adaptive sampling perspective, Bayesian inference is a better-suited form of inference for two major reasons. The first reason is that Bayesian inference is more naturally able to handle sequential sampling. The second reason is that Bayesian inference offers a far more generalisable approach to quantifying uncertainty thanks to methods such as marginalisation, Monte-Carlo integration and MCMC methods. In terms of the four stages of adaptive sampling, these advantages translate over to substantial advantages at the updating uncertainty stage. Furthermore, Bayesian inference synergises well with the other methods introduced later in this chapter (3.2.3-3.2.5). This is because many methods introduced in this section rely on estimated parameter values. Under Bayesian inference, uncertainty in these estimates can be accounted for via marginalisation and simulation methods. In general, it is difficult to provide a frequentist equivalent to this without relying on large sample sizes, which is unlikely to hold for early sub-samples.

Whilst ideally, one would want to be fully agnostic to the method of UQ (including the type of inference), the advantages Bayesian inference offers adaptive sampling are simply too beneficial to ignore when compared to frequentist inference. Consequently, this thesis will assume Bayesian inference in UQ from now. It should be noted though that there is still a wide variety of UQ methods under this restriction, and no further assumptions are required in this framework.

3.2.3 Using predictors in a model as a basis for targeted sampling

An important requirement for managing uncertainty effectively in any adaptive sampling framework is the ability to quantify uncertainty from targeted and biased sample designs. This is because adaptive sampling is typically motivated by situations where uncertainty needs to be balanced alongside design restrictions (e.g. limited sample sizes, less accessible members of the population etc). In these situations, an ability to target sampling toward different members of the population and make use of data collected under biased sampling greatly opens the options available when trying to efficiently manage these trade-offs.

With this need in mind, the thesis proposes that a simple way of quantifying uncertainty under targeted sample designs is to use the predictors in the model used as the basis for the said targeted sampling. This idea can also be easily extended to account for biased sampling (i.e., uncontrolled targeted sampling).

Whilst it is possible to use modelling to account for any form of targeted sampling in UQ [350], [351], a generalised approach for this may require heavy modelling assumptions. Supposing one was willing to accept such modelling assumptions, there may be an additional obstacle when it comes to fitting the models.

To understand why using the predictors to define targeted sampling makes quantifying uncertainty easier, consider how quantifying uncertainty can change under different design types in the general setting. With Bayesian inference, this can be seen by first conditioning Bayes theorem on a sample design S to give

$$\pi(\theta|D,S) \propto \pi(D|\theta,S)\pi(\theta|S).$$

Under this conditioning, the potential issues in targeted sampling can be then reframed as understanding how different sample designs S affect posterior distributions for θ. To deal with the conditional prior distribution, $\pi(\theta|S)$, first, one can assume θ is independent of S (i.e $\pi(\theta|S) = \pi(\theta)$) without any major implications or loss of generality. This is because the only way that this does not hold is if prior knowledge of θ is dependent on the sample design, which would border on absurd in practice, as it would require that prior knowledge in θ (i.e. before any data are collected) may change simply by proposing different sample designs with no need to implement them.

With the assumption that θ is independent of S, the issues from targeted (or biased) sampling in Bayesian inference ultimately comes down to how S affects the likelihood function, $\pi(D|\theta, S)$. One option here is to assume S away with independence, i.e $\pi(D|\theta, S) = \pi(D|\theta)$ for all S. This approach requires some heavy modelling assumptions though, which as discussed in Section 2.2.1, may be difficult to justify, especially when dealing with MLTs. An alternative option is to try and explicitly model $\pi(D|\theta, S)$. Assuming this is possible, one then has the problem of going from the likelihood function to the posterior distribution, $\pi(\theta|D, S)$. The problem here is not a mathematical one, as there are many methods in MCMC sampling that can handle bespoke likelihood functions in theory. Rather, the use of bespoke likelihood functions can create practical difficulties when it comes to their implementation. The reason for these difficulties is that many popular statistical software packages (which have been optimised and gone through a sufficient degree of quality assurance) implicitly assume sample designs where $\pi(D|\theta, S) = \pi(D|\theta)$ holds. Hence, if one were to rely on bespoke likelihood functions, one would need to go through substantial effort in building new (or editing existing) software packages. These problems are compounded in adaptive sampling as one has to go through such processes after every iteration.

One way of avoiding many of the issues that come from targeted sampling is to define targeted sampling in terms of the predictors in a model. This is because **any** sampling that can be defined using only the predictors in a model will stratify $\pi(D|\theta, S) = \pi(D|\theta)$. For a more formalised statement and proof see Box 3.5.

Using the predictors to define targeted sampling helps in the updating uncertainty stage of the adaptive sampling framework in three major ways:

1. It allows one to use third-party software (e.g. R packages including but not limited to mgcv [352], rjags [353], BayesGPfit [354], bmkr [355]) and results

related to conjugate priors [356] without any additional work as there is no need to adjust likelihood functions or customise pre-built methods for generating posterior distributions.

2. It is well suited for sequential sampling. This is because, if two probabilistic sample designs can be expressed as functions of a model's predictors, so can their composition. Note, one does not need to explicitly formulate these functions or their compositions. One only needs to know that the functions exist for the result to extend to sequential sampling.

3. The approach can be reverse-engineered to offer an easier way of accounting for sample bias. If one has a biased sample and a good idea of what is influencing this bias (e.g. distance, cost, age etc.), then this can be accounted for by including these features in a model from the start. This practice may be especially useful in the earlier iterations of adaptive sampling as one is typically given a biased sample rather than collecting data under controlled and targeted designs. Once again, one does not need to explicitly state how the bias is defined but only know that it is some function of the influencing factors. Note though, that one will typically need enough flexibility in their model (e.g. data-driven modelling) as justifying explicit model structure with the additional variables may not be easy. Furthermore, the factors influencing the bias may not have much predictive power but are nevertheless required in any model to use this method. This latter point is relevant as many model selection methods (e.g. Bayes factors) may falsely indicate that these features are not needed.

In addition to offering many advantages at the updating uncertainty stage, using the predictors to define the targeted sample combines well with other methods at the design proposal stage. The high-level idea here is to try to find clusters of desirable cases within the feature space of the predictive features (e.g. finding influential points or concertation of points with a high degree of epistemic uncertainty) and then define the targeted sampling across this feature space to focus on these cases. Under this approach, one can create efficient sample designs that are easy to apply under third-party statistical software packages and do not require any further modelling assumptions.

Box 3.5. *A formalised statement and proof for why using the predictors as a basis for targeted sampling can make accounting for bias and targeted sampling much easier.*

Claim: Let I denote a subset of a population and S denote some sample design. Next, let $D(I) = \big(Y(I), X(I)\big)$, where $Y(I)$, $X(I)$ denotes the outcomes and predictors under a model with parameters θ for the members of the population contained in I respectively. If there exists a g such that $\pi(I|S) = g(X)$ then $\pi(\theta|D,S) = \pi(\theta|D)$

Proof: let S be a sample design such that $\pi(I|S) = g(X)$. Next, one can consider two equivalent expressions for the joint probability distribution, $\pi(D,S|\theta)$

In the first case, one has

$$\pi(D,S|\theta) = \pi(S|D,\theta)\,\pi(D|\theta)$$

Which comes from the definition of a conditional distribution and is true for any S.

From this, one can make use of the condition that $\pi(I|S) = g(X)$. Here, $\pi(I|S) = g(X)$ implies that $\pi(S|D,\theta) = \pi(S|X)$ as if X is known, all other information redundant when determining the likelihood that the data were sampled under S. This gives one form to the joint probability distribution as

$$\pi(D,S|\theta) = \pi(S|X)\,\pi(D|\theta). \tag{1}$$

The second form for $\pi(D,S|\theta)$ can be given by fist conditioning on the S to give

$$\pi(D,S|\theta) = \pi(D|\theta,S)\pi(S|\theta).$$

Since the design of S is determined only by X, the likelihood of S is unaffected by θ. This gives $\pi(S|\theta) = \pi(S)$. Hence the second expression for $\pi(D,S|\theta)$ becomes

$$\pi(D,S|\theta) = \pi(D|\theta,S)\pi(S). \tag{2}$$

Comparing the right-hand sides of (1) and (2) gives

$$\pi(D|\theta,S) = \frac{\pi(S|X)}{\pi(S)}\pi(D|\theta) \propto \pi(D|\theta). \tag{3}$$

The desired result, $\pi(\theta|D,S) = \pi(\theta|D)$, follows immediately form (3), as under a fixed prior distribution $\pi(\theta)$, $\pi(D|\theta,S) \propto \pi(D|\theta) \Rightarrow \pi(\theta|D,S) = \pi(\theta|D)$.

3.2.4 Quantifying aleatoric and epistemic components of uncertainty

As introduced in Chapter 1 and discussed in Section 2.1, the aleatoric components and epistemic components play an important role in managing uncertainty efficiently. This is because the aleatoric and epistemic components of uncertainty are useful measures to indicate when uncertainty in predictions can be reduced with further sampling and when one is approaching the limits of what sampling alone can do in reducing uncertainty.

When trying to quantify the aleatoric and epistemic components of uncertainty though, these components are themselves subject to uncertainty as they often rely on unknown parameter values. Under Bayesian inference, it is possible to account for this uncertainty through marginalisation (along with Monte Carlo integration if need be). To see why this is the case, one can consider an example where the precision of predictions and estimates are measured using variances.

Suppose one has a model $y = f(x;\theta)$ where x is a vector of predictors and θ denotes a set of parameters. Here the *aleatoric variance* for y under f is given by

$$a_V(y;f) = V(f(x)).$$

The value a_V is a way of quantifying the maximum level of precision for y under f when the precision of an estimate or prediction is measured by its variance. Note, a_V does not mean estimates for y could not be made more precise with further modelling, nor does it deal with any ontological uncertainty that comes from the assumptions in the model f.

Since a_V will often depend on the unknown parameter values, a_V will need to be estimated and so will be subject to uncertainty. With Bayesian inference, this uncertainty can be accounted for by marginalising over the parameter values to give a posterior distribution for a_V. That is

$$a_V(y;f)|D = \int \pi\big(V\big(f(x)\big)\big|\theta\big)\pi(\theta|D)\,d(\theta).$$

Assuming no closed-form expression for $a_V(y; f)|\ D$ can be found, one can use Monte Carlo integration to generate $a_V(y; f)|D$.

From $a_V(y; f)|D$ the *epistemic variance* for y under f given observed data D can be determined with

$$e_V(y; f\ )|D = V(y; f|D) - a_V(y; f)|D,$$

where $V(y; f|D)$ denotes the variance of predictive distribution for y under f given observed data D.

With this example, one can see that Bayesian inference is a highly generalisable approach to accounting for uncertainty when quantifying aleatoric and epistemic components of uncertainty, as providing that one can sample from the posterior distribution for the parameter values, $\theta|D$, one can easily propagate uncertainty in the parameters when estimating aleatoric and epistemic uncertainty.

As a side note, the principles of aleatoric and epistemic variance can be extended to include other forms of uncertainty measures (e.g. the length of credible intervals). However, this thesis will stick to precision measures based on variances, as linear properties of the variances tend to make the calculations more mathematically convenient.

Quantifying aleatoric and epistemic components of uncertainty is useful in the design proposal stage of an adaptive sampling framework as they indicate which kinds of predictions are likely to benefit from further sampling and when other kinds of predictions are close to their maximum precision under a model. This information can then be leveraged to inform targeted sampling for future iterations. Quantifying aleatoric components of uncertainty can also play an important role in the design assessment stage of adaptive sampling, as it can help contextualise any analysis by giving users an idea of the maximum effectiveness of any sample design under a particular model.

3.2.5 Predicting the likely effects of further sampling

Under any adaptive sampling procedure, there will become a point where one will have to decide how to continue with further sampling. This decision could involve how to target sampling towards different members of a population, the total sample size, or even the decision to end the adaptive sampling procedure altogether.

One key lesson from the literature review is that the effects of these decisions are themselves subject to uncertainty. This uncertainty can come from uncertainty in parameter values, uncertainty in model choice and the stochastic nature of many sample designs.

Unfortunately, this uncertainty limits the utility of past studies and post-hoc analysis when making sampling decisions. Figure 3.6 provides an overview of the major challenges one faces when trying to use past studies to justify sample designs in new situations. Briefly, these major challenges are:

(i) Deciding when previous problems are similar enough to be relevant to a new case.

(ii) Using past case studies for justification of sample designs does not hold when dealing with a novel case.

(iii) Past studies may not consider the same sets of methods, making like-for-like comparisons difficult.

(iv) Uncertainty in key parameter values can make it difficult to judge if a new case is similar enough to past studies.

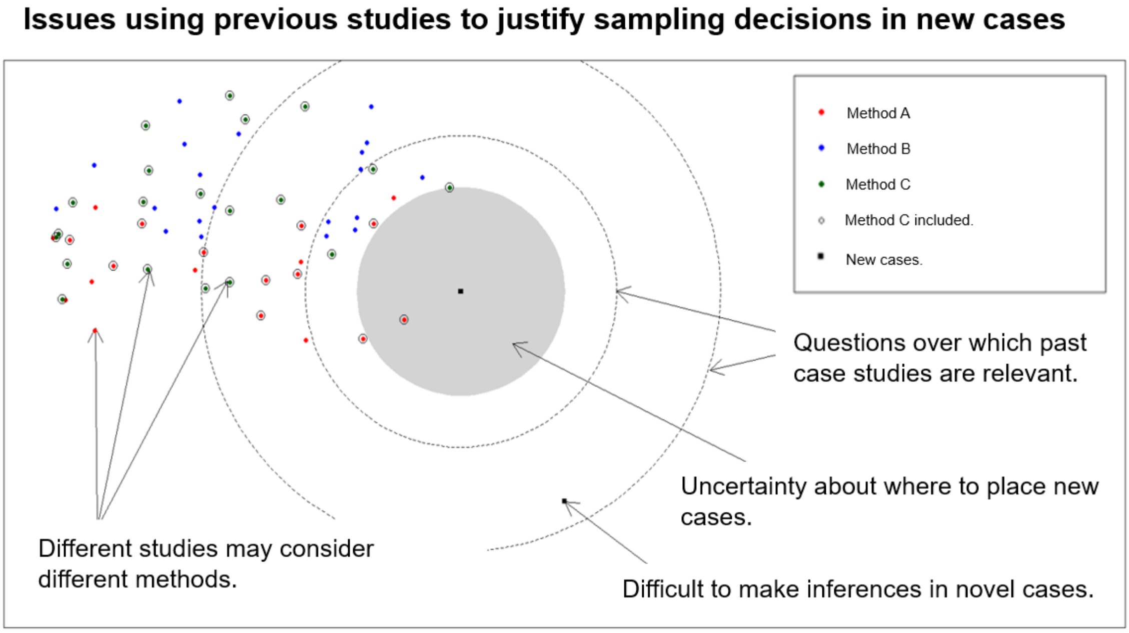

Issues using previous studies to justify sampling decisions in new cases
Method A
Method B
Method C
Method C included.
New cases.
Questions over which past case studies are relevant.
Uncertainty about where to place new cases.
Different studies may consider different methods.
Difficult to make inferences in novel cases.

Figure 3.6. *An illustration of the challenges one faces when using past studies or simulation studies to justify efficient sample designs. This example considers three methods for proposing efficient sample designs (Methods A, B and C). The coloured points represent past studies with the colours representing the most efficient method in that study (e.g. a red point represents that method A was the most efficient in that study). The black squares represent new cases. The encircled points represent case studies where method C was not considered.*

Because of these challenges, this thesis proposes that sampling decisions should be based on predicting the likely effects each design will have on reducing uncertainty using probability distributions. Figure 3.7 illustrates the sort of outputs one would expect under this alternative approach.

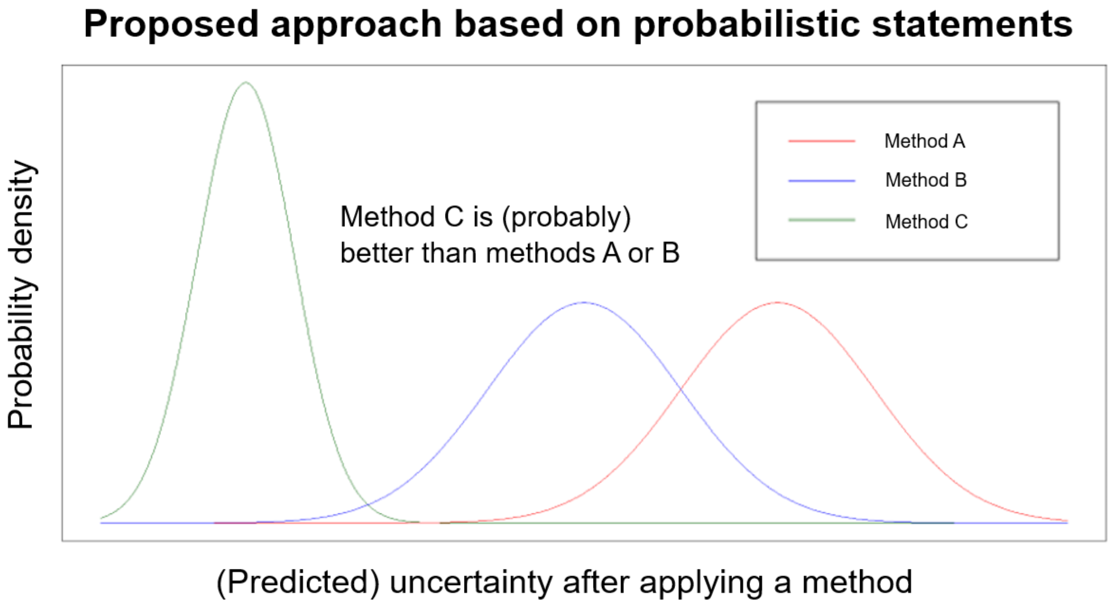


Figure 3.7 *An illustration of how one would use probability distribution functions to predict (then compare) the different methods of generating sample designs. Here, one uses the current information within the new case to make probabilistic statements regarding sample designs that have not been implemented yet. In this example, one would conclude Method C is likely to be the better option based on current understanding.*

Fortunately, generating these probability distributions using the current information is relatively easy with Bayesian inference. Effectively, predicting the likely effect of further sampling under Bayesian inference is simply two applications of marginalisation. As an example, suppose one wishes to assess how an iteration of sample design S may affect the variance in the prediction of y under a model f after observing data D. Using marginalisation, one has

$$V(y, f|D, S)|D = \iint \pi\big(V(y, f|D, D^*)\big)\pi(D^*|\theta, S)\, \pi(\theta|D)\, dD^* d\theta,$$

where $D^*(S)$ is a set of data obtained under design S (note D^* is usually a random variable as S typically has a stochastic component).
In practice, an expression like this may be difficult to solve analytically. Thankfully though, one can generate a sample form $V(y, f|D, S)|D$ through Monte Carlo methods. More specifically one can apply the following steps:

- Step 1, generate θ^* by drawing a sample from $\theta|D$.
- Step 2, generate an artificial sample D^* by simulating S under the assumption that $\theta = \theta^*$.
- Step 3, calculate $V(y, f|D, D^*)$.

This procedure can be repeated multiple times to build a distribution for $V(y; f|D, S)|D$.

Predicting the likely effects of further sampling plays a major role in the design assessment phase of the adaptive sampling framework as it gives users an evidence-based method of comparing different sample designs before they are implemented. In these cases, answers are given in probabilistic terms. For example, questions such as *“how large does a reference sample need to be to reach a set level of precision?”* are answered with a statement such as *“based on the current data, one can expect a sample under design A with a sample size of between M and N to meet this goal. However, if one were to consider a targeted sample design under B, this sample size may only need a sample size between K and L”.*

3.3 Example Workflow

Suppose one has a model $y = f(x, \theta)$ where x, θ are the predictors and parameters in model f respectively.
Next suppose that one wishes to reduce uncertainty in y using data of the form $D = (Y, X)$ under a measure of uncertainty $U(y, D)$ (e.g. $U(y, D) = v(y|D)$, the variance of the posterior distribution for y given D).

Table 3.8 provides a worked example for using the proposed adaptive sampling framework to suggest sample designs that efficiently reduce uncertainty in y for different values of x.

Whilst left at a more general level, the example workflow in Table 3.8 can be used for uncertainty management in MLTs. Here, there are two scenarios this workflow may be applied to. The first is when f is a MLT and the x values are some predictors. The second scenario is when the x values are the outputs of a MLT and f is an intermediary model for UQ purposes. Note, the benefits and drawbacks of these different model types have already been discussed in Chapter 2 and this comment is merely a statement that the workflow **can** be applied in both cases.

Table 3.8 (Part I) *A generalised workflow for the adaptive sampling framework (and methods introduced throughout Section 3) alongside a worked example.*

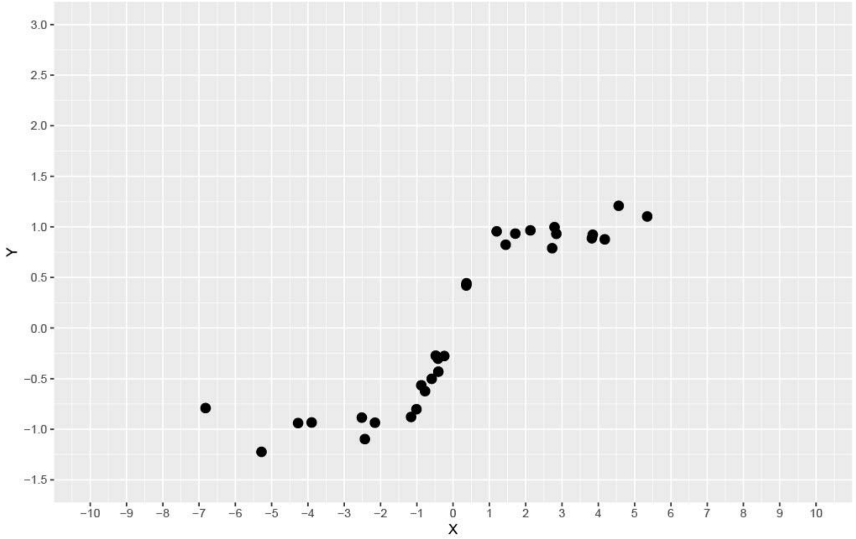
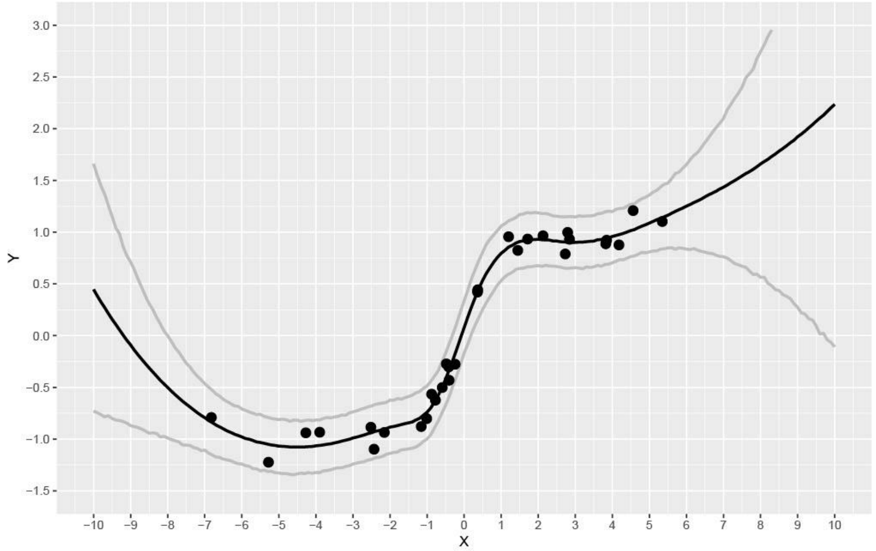
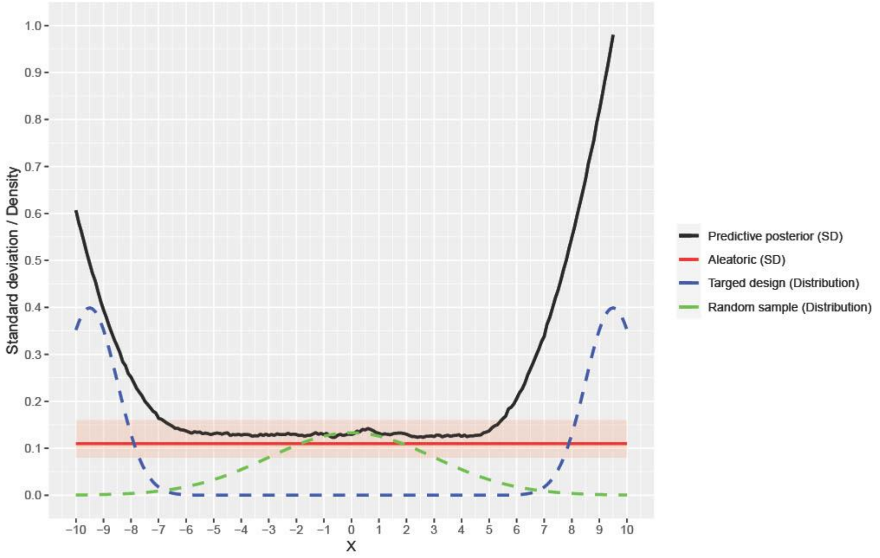
Stage	Core	Optional
Step 1 *(Updating the sample)*	Obtain an initial sample D under a design S that satisfies $\pi(I\|S) = g(X)$.	If an initial sample design does not satisfy this condition but does satisfy $\pi(I\|S) = h(W)$, for some set of features W, then an alternative model of the form $y = f^*((x,w);\theta^*)$ can be made to satisfy this condition.
Example *The initial data are collected under a simple random sampling. Simple random sampling satisfies the condition with $\pi(D\|S) = \pi(X)$.*		
Step 2 *(Updating Uncertainty)*	Generate the posterior distribution $\mathrm{y}\|D$.	Use marginalisation and Monte Carlo methods to generate $\mathrm{y}\|D$ from $\theta\|D$
Example *$y = f(x;\theta)$ is based on a generalised additive model* [357] *$U(y, D)$ is the standard deviation of the posterior distribution of y. i.e.* $U(y, D) := \sqrt{v(y\|D)}$		

Table 3.8 (Part II) *A generalised workflow for the adaptive sampling framework (and methods introduced throughout Section 3) alongside a worked example.*

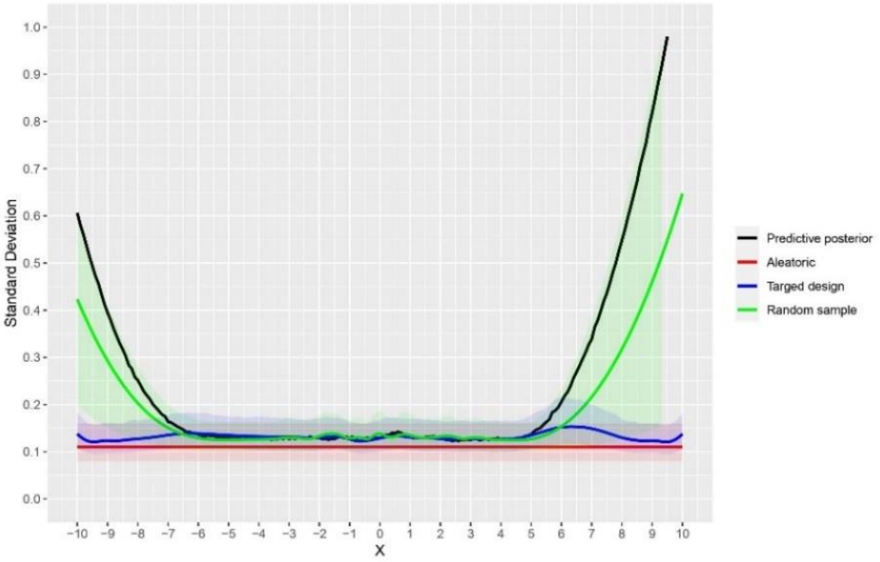
<table>
<tr><th>Stage</th><th>Core</th><th>Optional</th></tr>
<tr><td><b>Step 3</b>
<i>(Design proposal)</i></td><td>Propose sample designs $S_1, \ldots., S_n$. Such that $\pi(I|S_i) = g_i(X)$</td><td>Estimate aleatoric and epistemic components of z^* to help generate proposal designs.</td></tr>
<tr><td><b>Example</b>

<i>Two proposal designs, both of size 30.</i>

<b><i>Blue:</i></b> <i>Targeted sampling towards areas with substantial epistemic uncertainty.</i>
<b><i>Green:</i></b> <i>Simple random sampling (again).</i></td><td colspan="2">


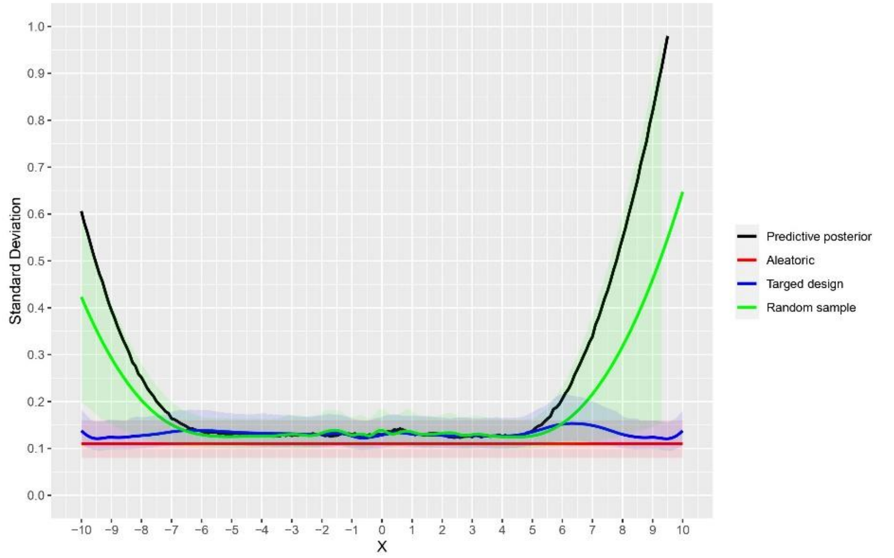
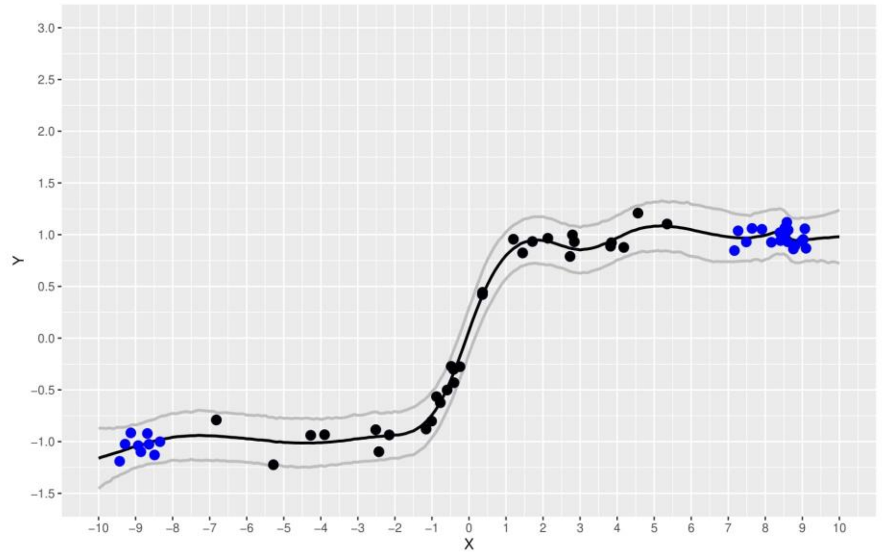
</td></tr>
<tr><td><b>Step 4</b>
<i>(Design assessment)</i></td><td>Estimate the likely effects of sample designs $S_1, \ldots., S_n$ will have on the measure of uncertainty by generating posterior distributions

$$U(y, \{D, D_i^*\})|D$$

Where D_i^* denotes data obtained under a sample design S_i</td><td>Use marginalisation and Monte Carlo methods to generate the posterior distributions from $\theta|D$</td></tr>
<tr><td><b>Example</b></td><td colspan="2">


</td></tr>
</table>

Table 3.8 (Part III) *A generalised workflow for the adaptive sampling framework (and methods introduced throughout Section 3) alongside a worked example.*

Stage	Core	Optional
Step 5 (Design assessment)	Decide upon a design S^* from $S_1, \dots., S_n$ or select no further sampling. If no further sampling is selected end here.	Consider the aleatoric component of uncertainty to help decide if no further sampling should be selected and assess sample designs.
Example *The targeted sample design appears to be better at reducing uncertainty in the areas of substantial epistemic uncertainty.*		
Step 6 *(Updating the sample and updating uncertainty)*	Implement S^* to obtain data D^* and return to step 2 with $D \leftarrow \{D, D^*\}$	
Example		

Additional comments

From the example in Table 3.7, there are three noteworthy observations to comment on. Firstly, there is an option to use marginalisation and Monte Carlo methods in many of the steps. Whilst strictly not necessary (e.g., one could find closed-form solutions to posterior distributions under the right conditions), the fact that this is an option under Bayesian inference greatly improves the generalisability of the proposed framework.

Secondly, a case could be made that the condition in step 3 $\left(\pi(I|S_i) = g_i(X)\right)$ is optional, as one could use other sample designs if they were willing to make enough modelling assumptions. However, removing this condition will make steps 4 and 6 noticeably more difficult.

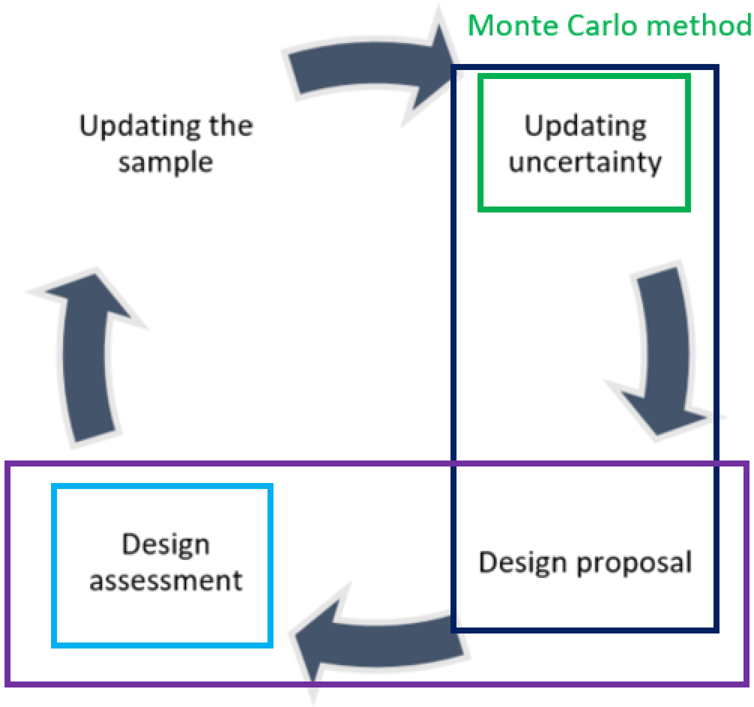
Thirdly, from all the practices discussed in sections 3.2.2-3.2.5, estimating different components of uncertainty is the most optional when creating sample designs. For example, one could use any means of proposing sample designs with optimisation methods. However, estimating the aleatoric component of uncertainty is recommended for at least step 5, as this will help a user in deciding when to stop sampling.

3.4 Summary

When discussing ideas related to how uncertainty from machine learning classifiers may be better managed, one must consider that the specific MLTs and methods of UQ may be quite different across scenarios. Hence, this chapter introduced a framework for uncertainty management based on adaptive sampling that aims to be agnostic to the choice of UQ and MLT.

To achieve this framework, this chapter began by breaking adaptive sampling into a cyclical process involving the four key phases: updating the sample, updating uncertainty, design proposal, and design assessment. Following this, the chapter then populated this framework with a number of methods and proposed practices. These methods and practices were using Bayesian inference in UQ; using predictors in a model as a basis for targeted sampling; quantifying aleatoric components of uncertainty; and predicting the likely effects of further sampling. A summary of how these methods interact with the stages of adaptive sampling is provided in Figure 3.8.

Bayesian inference makes updating uncertainty under sequential sampling much easier and synergies well with Monte Carlo methods.



Predicting the likely effects of further sampling is a highly generalisable method in the Design assessment phase that is useful when deciding between a short list of potential designs in further iterations.

Using the predictors in a model to define targeted sampling at the Design proposal phase makes it easier to update the uncertainty in subsequent iterations of sampling.

Quantifying aleatoric components of uncertainty acts as a way of identifying which areas to target in further iterations in the Design proposal stage and helps in the Design assessment phase by contextualising any analysis.

Figure 3.8. *A summary of how the methods proposed in Section 3.2 interact with the four key stages of adaptive sampling.*

With this cyclical process and set of methods, a framework for adaptive sampling has been set. Under this framework, the methods are designed to address different challenges one would expect to see in adaptive sampling. Briefly, the intended contribution of each method is as follows:

- Using Bayesian inference in UQ lays the foundation for all remaining methods as it allows uncertainty in parameter values to be propagated using simulation-based methods. It also enables the updating of uncertainty between iterations, as it is naturally suited to sequential sampling.

- By using the predictors in a model as a basis for targeted sampling, one can pre-emptively ensure that updating the uncertainty after the next iteration of sampling is a smooth one and avoid many complications that can arise when using targeted sampling in a general setting. This idea can be extended to include non-deliberate sample bias by having the factors that influence the bias in the model from the beginning.

- Quantifying aleatoric and epistemic components of uncertainty helps identify which members should be targeted to manage uncertainty efficiently and indicate when it is best to stop sampling under a fixed model structure.

- Predicting the likely effects of further sampling offers users a consistent way of confirming and assessing proposed sample designs without implementing them. This can be used to make informed sampling decisions for future iterations and experiment with different "what if?" scenarios.

With the proposed framework for adaptive sampling established, the next stage is to investigate how this framework meets the assessment criteria in practice under the two case studies.

Chapter 4 Case Study 1: Lagos Urban Mapping

4.1 Problem introduction

This chapter evaluates the adaptive sampling framework introduced in Chapter 3 using the Lagos urban mapping problem first introduced in Chapter 1.

To recall from Section 1.3, the Lagos case study focuses on managing uncertainty around the Lagos area in an urbanisation mapping problem. This first case study provides two scenarios in which to evaluate the adaptive sampling framework. The first scenario considers an area estimation problem using the discrete 30m maps (part I), whilst the second scenario involves managing uncertainty at the pixel level for a 1km resolution map that uses fuzzy classifications (part II).

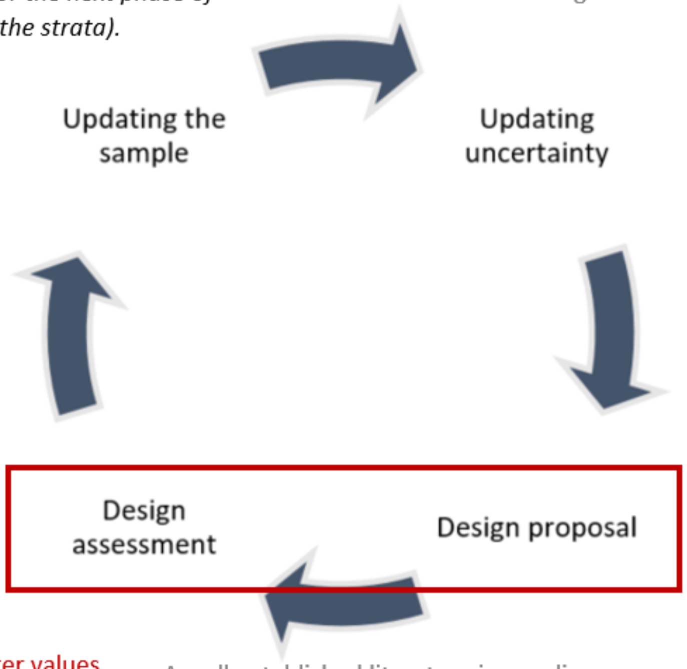
The two parts of this Lagos study are intended to act as first-level examples before building up to a second case study. More specifically, the Lagos urbanisation mapping problem acts as a real-world case study where one has access to a full reference map. Having a full reference map is useful at this stage of the analysis, as it gives a space where adaptive sampling practices can be tried out and refined without needing to worry about the practical restrictions of sampling (although various hypothetical sampling limitations will be proposed as part of the evaluation). Naturally, unrestricted access to all areas and a full reference map will not be available in genuine applications, and this will be considered more explicitly during the second case study in Chapter 5.

From the perspective of the adaptive sampling framework, part I investigates the utility of Bayesian inference when predicting the likely effects of different stratified random sample designs on aggregate-level estimates, e.g. class prevalence, overall accuracy, sensitivity, specificity etc. Because of the relative simplicity in the method of UQ (resulting in little ontological uncertainty and no aleatoric uncertainty) and well-established literature on generating efficient sample designs under stratified random sampling, one can easily move past the updating uncertainty and design proposal stages in the framework and move straight to the design assessment phases (see Figure 4.1 for a summary).

Start: *the initial sample has been collected under stratified random sampling.*

Objective: *decide upon an appropriate design for the next phase of sampling (total sample size, distribution across the strata).*

UQ for total urbanisation is based on the sum of beta distributions. Each distribution is generated using binominal distributions. No major issues.



Uncertainty in parameter values makes proposing and assessing sample designs more difficult.

A well-established literature in non-linear programing to help suggest optimal sample distributions under fixed sample sizes.

No aleatoric component of uncertainty.

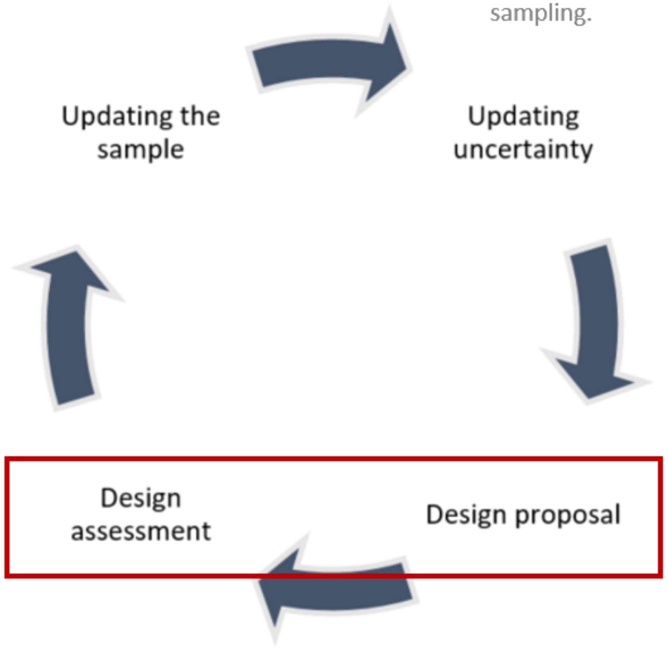
Figure 4.1. *A summary of how the first part of the Lagos urban mapping study relates to challenges across the four key stages of adaptive sampling.*

The second part of the Lagos case study builds upon part I with consideration of UQ for individual instances which brings additional challenges. Firstly, there is now an aleatoric component of uncertainty in the model; secondly, the model lacks many closed-form solutions compared to part I, forcing one to use Monte Carlo methods when proposing and assessing sample designs (see Figure 4.2 for a summary).

Start: *the initial sample has been collected under targeted sampling defined through predicted urban extent.*

Objective: *decide upon an appropriate design for the next phase of sampling (total sample size, how to target sampling).*

UQ is based on modelling reference extents from predicted urban extents. The model accounts for heterogeneous errors and truncated values. Since the strata are defined using the predicted extents, there are no issues due to targeted sampling.



Uncertainty in parameter values makes proposing and assessing sample designs more difficult.

Methods for suggesting optimal balancing of trade-offs are not readily available.

Aleatoric uncertainty exists and limits the maximum level of precision in modelling.

Figure 4.2*. A summary of how the second part of the Lagos urban mapping study relates to challenges across the four key stages of adaptive sampling.*

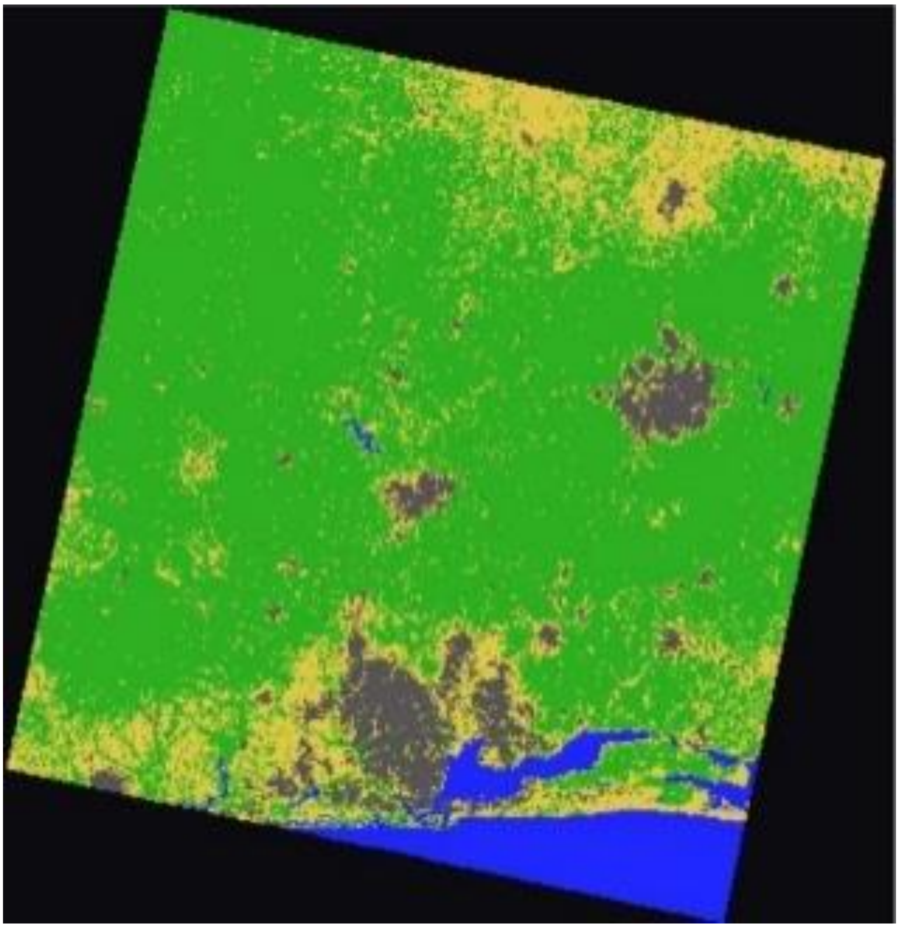
4.2 Lagos urban mapping (I): estimating the urban area from a discrete classifier

The first part of this case study begins with estimating the total urbanised area using the 30m resolution map. Estimating the degree of urbanisation at a given time can be a useful predictor when monitoring population growth [358], [359] or carbon emissions due to land-use change [360], [361], [362]. If the prediction map were a perfect classifier, the degree of urbanisation in this scenario would be proportional to the number of areas classified as urban (note, this scenario puts aside any ontological uncertainty related to whether areas at a 30m resolution are realistically fully urban or fully non-urban). However, the possibility of misclassifications creates a degree of uncertainty in any area estimations.

To quantify uncertainty in the area estimates due to potential misclassifications, this first scenario begins with a stratified random sampling of 1000 pixels where one can compare predicted classifications with the reference values. The initial sample size of 1000 was not based on any objective criteria. However, this choice was motivated by the following:

- The initial estimate should be precise enough that adaptive sampling is feasible. If the initial sample size is too low, one cannot get enough meaningful information to apply adaptive sampling effectively.
- The initial estimate should not be so precise that it negates the need for any further sampling.
- There should be a reasonable amount of reference data from each stratum after using the initial sample design.

The four strata are selected based on the predicted map, and their sample sizes are proportional to the relative spatial area of each stratum. The first two strata are the predicted classes of *water* and *urban land*. The remaining pixels (all of which belong to the predicted non-urban land stratum) are divided into two strata based on which may be prone to errors. A pixel is decided as being prone to error if any of the decision trees from the original Random Forest classifier predicted the pixel as urban. The motivation for this splitting comes from the work discussed in Chapter 2 where efficient sample designs are more easily generated when one first divides the area into smaller noisy strata and large homogenous strata. Figure 4.3 displays the strata map and confusion matrix from the initial sample.



		Reference Data				Stratum sizes	
	Predicted Class	**Urban Land**	**Water**	**Nonurban Land**	**Total**	**Proportion (W_i)**	**Absolute (Km^2)**
⬛	*Urban Land*	56	0	7	63	0.063	2303.7
🟦	*Water*	0	60	1	61	0.062	2267.1
🟨	*Nonurban Land 1*	10	0	163	173	0.173	6325.9
🟩	*Nonurban Land 2*	1	0	702	703	0.703	25706.0

Figure 4.3. *An urban mapping of the Lagos area in 2016 based on discretely classified pixels at a 30m resolution along with a confusion matrix from an initial reference sample. Key: Urban land (grey), Water (blue), and Nonurban Land (yellow and green). Yellow Nonurban Lands indicate areas within the nonurban area that are suspected of being more prone to errors.*

With this initial sample, the uncertainty for the proportion of urbanised area U is quantified using A with

$$A = \sum_{i=1}^{4} W_i A_i$$

Where

$A_i \sim beta\left(x_i + \frac{1}{2}, n_i - x_i + \frac{1}{2}\right),$

n_i is the total number of pixels randomly selected from stratum i,

x_i denotes the number of sampled pixels that are categorised as urban by the reference map from stratum i,

W_i is the relative size of stratum i with $\sum_{i=1}^{4} W_i = 1$.

Under A, uncertainty for U may be quantified under a Bayesian or frequentist perspective. From a Bayesian perspective, A is a posterior distribution based on a weighted sum of within-strata urban proportions, A_i, which have all been assigned Jeffreys prior distributions. Under a frequentist perspective, A may be viewed as an extension of Jeffreys intervals, where the percentiles of A may be used to construct confidence intervals (e.g. a 90% confidence interval may be made by taking the 5th and 95th percentile of A). In either case, an estimation for the total urbanised area may be given with $\widehat{U} = \sum_{i=1}^{4} W_i \frac{x_i + \frac{1}{2}}{n_i + \frac{1}{2}}$ and the precision of this estimate may be measured with the standard deviation of A, i.e $\sqrt{V(A)}$. The motivation for this measure is that $\left(\sum_{i=1}^{4} W_i \frac{x_i + \frac{1}{2}}{n_i + \frac{1}{2}}\right) \pm \left(z_{1-\frac{\alpha}{2}} \times \sqrt{V(A)}\right)$ represents an approximate $100(1-\alpha)\%$ credible (or confidence) interval.

Under the initial sample, the estimated urban proportion is 0.0685 (2489 km^2) with a precision of 0.0042 (155 km^2), To put this precision into context, 155 km^2 is approximately 6.3% of 2489 km^2 and annual growth rates in urbanised areas from 1990 to 2000 at national levels varied from around 2.9% to 7.2% depending on that nation's level of development and across different continents [313], [314]. Hence a precision of around 6.3% in this context justifies a need for further sampling, especially if one wants to estimate the overall growth of urbanised land over a few years.

In terms of uncertainty quantification, the model is simple. Any issues related to ontological uncertainty are going to be minor, as it is based on weighted sums of proportions that are each estimated based on binomial distributions. In addition, there is no need to consider aleatoric uncertainty in this case, as it is known that there is no aleatoric component under A. The use of beta distributions for each A_i allows for an easy way of updating uncertainty, especially when under a Bayesian setting due to conjugacy [356].

One joins this first part of the case study at the design assessment phase of the adaptive sampling framework. Here, four sample designs are proposed based on stratified random sampling (see Table 4.4 for further details).

Table 4.4. *Confusion matrix from the initial sample along with the proposed sample designs (i)-(iv). The values in the proposed sample designs represent the number of pixels set to be randomly selected from each of the four strata (e.g. design (i) collects 63 pixels from the Urban land stratum, 61 pixels from Water stratum, etc).*

		Reference Data				**Proposed sample designs**			
	Predicted Class	**Urban Land**	**Water**	**Nonurban Land**	**Total**	**(i)**	**(ii)**	**(iii)**	**(iv)**
⬛	*Urban Land (1)*	56	0	7	63	63	350	250	189
🟦	*Water (2)*	0	60	1	61	61	0	100	183
🟨	*Nonurban Land 1 (3)*	10	0	163	173	173	650	500	519
🟩	*Nonurban Land 2 (4)*	1	0	702	703	703	0	100	2109

The motivation for the choice of each sample design is as follows:

- Design (i) is simply the initial sample design implemented once more (i.e. another 1000 pixels sampled where the sample sizes for each stratum are proportional to W_i). This is included to act as a benchmark as to what would happen if one were to continue with the same sample design.

- Design (ii) is based on using non-linear programming to suggest a design that optimally minimises $\sqrt{V(A)}$ supposing that a further 1000 pixels could be sampled and that the cost of sampling within each stratum is the same (see Chapter 2.2.2 for work related to using non-linear programming to suggest optimal sample designs). Here, design (ii) suggests that sampling should focus resources on the nonurban land ($i = 3$) and predicted urban ($i = 1$) strata, with the majority of the resources devoted to the former.

- Design (iii) is once again based on a total sample size of 1000. This time though, design (iii) attempts to acknowledge the suggested distribution in design (ii), but also balances the distribution to consider that one may need to use the map to estimate values other than the level of urbanisation. For example, user, producer, and overall accuracies are performance metrics that are often used to assess the quality of land cover maps [365], [366].

- Design (iv) is another benchmarking sample design. (iv) is proportional to sample design (i) but collects an additional 3000 pixels. Given that sample designs (i-iii) involve a total of 2000 (1000 in the initial sample plus a further 1000), design (iv) represents the initial sample with a total sample size double these previous three designs (i.e. 1000 + 3000 to give a total of 4000 pixels).

The fundamental problem in this first scenario is deciding which one of these sample designs (if any) should be implemented in the next phase of sampling. Fortunately, the effect any sample design will have on the uncertainty of the total urbanised area can be explicitly formulated in this situation. Firstly, let $A(j)$ denote update of A under sample designs $j = 1, \dots, 4$ that is

$$A(j) = \sum_{i=1}^{4} W_i A_i(j)$$

with

$$A_i(j) \sim Beta\left(x_i + x_{i,j}^{*} + \frac{1}{2}, n_i + n_{i,j} - (x_i + x_{i,j}^{*}) + \frac{1}{2}\right)$$

(4.1)

Where $n_{i,j}$ denotes the number of pixels drawn from stratum i under design j, $x_{i,j}^{*}$ denotes the number of pixels obtained under design j within stratum i that are labelled as urban areas by the reference map.

Next, using the fact that for $X \sim Beta(\alpha, \beta)$ its variance can be written as $V(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$, one can explicitly write down the effect each design will have on the precision measurement of the total urbanised area with

$$\sqrt{V\big(A(j)\big)} = \sqrt{\sum_{i=1}^{4} \frac{W_i^2 \left(x_i + x_{i,j}^* + \frac{1}{2}\right)\left(n_i + n_{i,j} - (x_i + x_{i,j}^*) + \frac{1}{2}\right)}{\left(n_i + n_{i,j} + 1\right)^2 (n_i + n_{i,j} + 2)}}$$

(4.2)

Furthermore, because $x_{i,j}^*$ is determined through random sampling within each stratum, each $x_{i,j}^*$ can be written as a binomial distribution with

$$x_{i,j}^* \sim Bin\big(p_i, n_{i,j}\ \big)$$

(4.3)

Where p_i denotes the proportion of urban pixels within stratum i that are classified as urban according to the reference map.

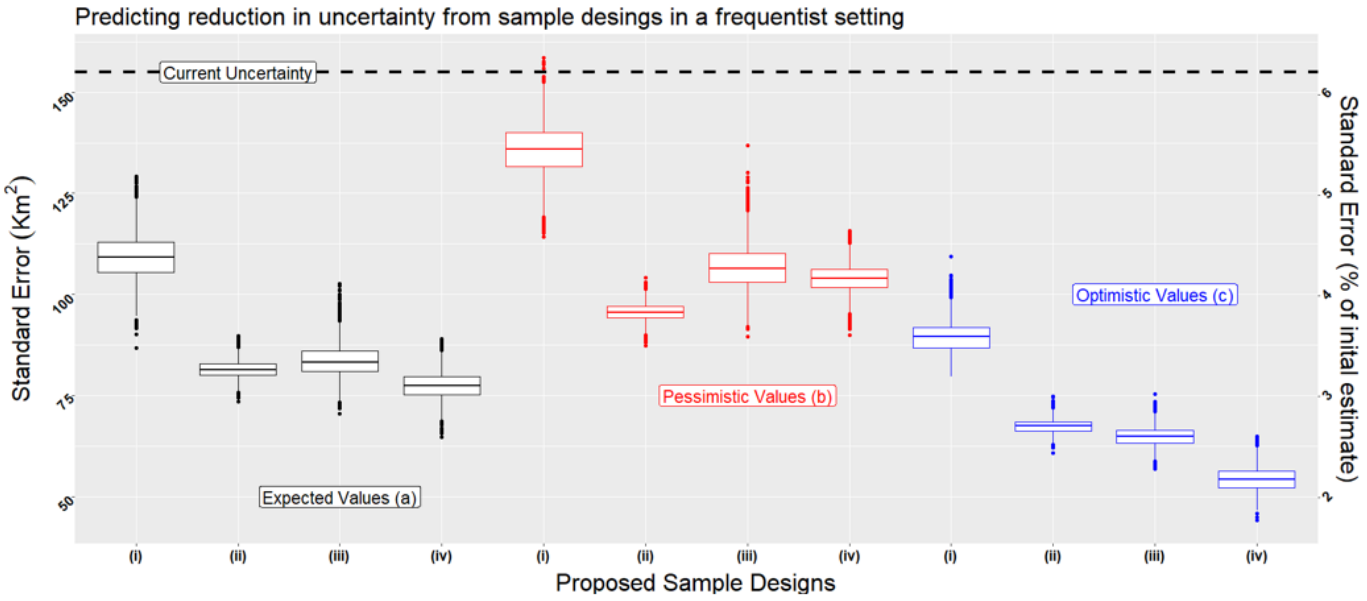
With this notation, the uncertainty in the total urbanised area under each sample design is set by a stochastic process influenced by a fixed p_i and variable $n_{i,j}$. However, since each p_i would be unknown in a real setting, there is also a degree of uncertainty within the value of each p_i.

It is at this point that one can investigate the utility of Bayesian inference when predicting the likely effects of further sampling using Bayesian inference. In terms of the notation and context set out so far, the underlying problem lies in how future sampling based on stratified random sampling will affect the precision measure, $\sqrt{V\big(A(j)\big)}$, when there is uncertainty in each p_i and a stochastic component for each $x_{i,j}^*$.

Predicting the effects of further sampling under a frequentist setting.

Under a frequentist setting, the uncertainty in each p_i (given the initial sample) may be quantified using confidence intervals constructed using the percentiles of each A_i (which equates to a Jeffreys interval in this scenario). Alongside this, one can predict the impact each design will have on the precision measure, $\sqrt{V\big(A(j)\big)}$, for a fixed set of p_i values via simulation methods based on (4.3). Combining these can offer a way of predicting the likely values for $\sqrt{V\big(A(j)\big)}$ by using the confidence intervals to give a plausible range of p_i values in a sensitivity analysis.

Applying this approach to the Lagos example, one can use the initial sample to construct 99% confidence intervals for each p_i (see Figure 4.5). Assuming the intervals represent a plausible range of p_i values, one may begin to compare the proposed sample designs across values within this four-dimensional cuboid. Figure 4.5 displays the results of three such plausible values. The first set of plausible values (a) assumes $p_i = \mathrm{E}(A_i)$ whilst sets (b) and (c) are based on the tails of the individual 99% confidence intervals designed to give pessimistic and optimistic estimates for $\sqrt{V\big(A(j)\big)}$ respectively.



A) Box plots of simulated precision measures across designs (i) –(iv)

		Proportion of Urban Pixels (99% Conf. Intervals)		
		Estimate	Lower bound	Upper bound
⬛	*Urban Land*	0.88281[a]	0.76197[b]	0.96362[c]
🟦	*Water*	0.00806[a]	0.00000[c]	0.07049[b]
🟨	*Nonurban Land 1*	0.06034[a]	0.02123[c]	0.11506[b]
🟩	*Nonurban Land 2*	0.00213[a]	0.00000[c]	0.00804[b]

B) The assumed proportions of urban pixels in each stratum

Figure 4.5 *(A) Box plots for the precision of the estimated urbanised area based on* 1×10^5 *simulations across the proposed sampling distributions using assumed proportion rates in (B) [black = (a), red =(b), blue = (c)]. (B) The assumed proportion rates in (a)-(c). Lower and upper bounds are based on confidence intervals at the 99% level.*

It is here that a major weakness of UQ under frequentist inference becomes apparent. Namely, there is no formal way of distinguishing between the results in Figure 4.5. This stems from the problem that under a frequentist setting, there is no formal way to distinguish between the values assumed in (a) - (c) as behind each confidence interval, there is only a statement related to whether the confidence interval will contain the true value of p_i in relation to repeated sampling. These statements make no claims related to whether the edges of these intervals are more or less plausible to those in the middle (or anywhere else in the interval for that matter). Hence, the outcomes of each set are equally plausible within the logic of confidence intervals.

To further compound this weakness, there is always going to be a degree of arbitrariness when deciding which values of p_i to consider, as there is no objective method for setting the level of confidence within any interval (or region in higher dimensional spaces). For example, one could have instead considered a similar methodology with 95% or 90% confidence levels.

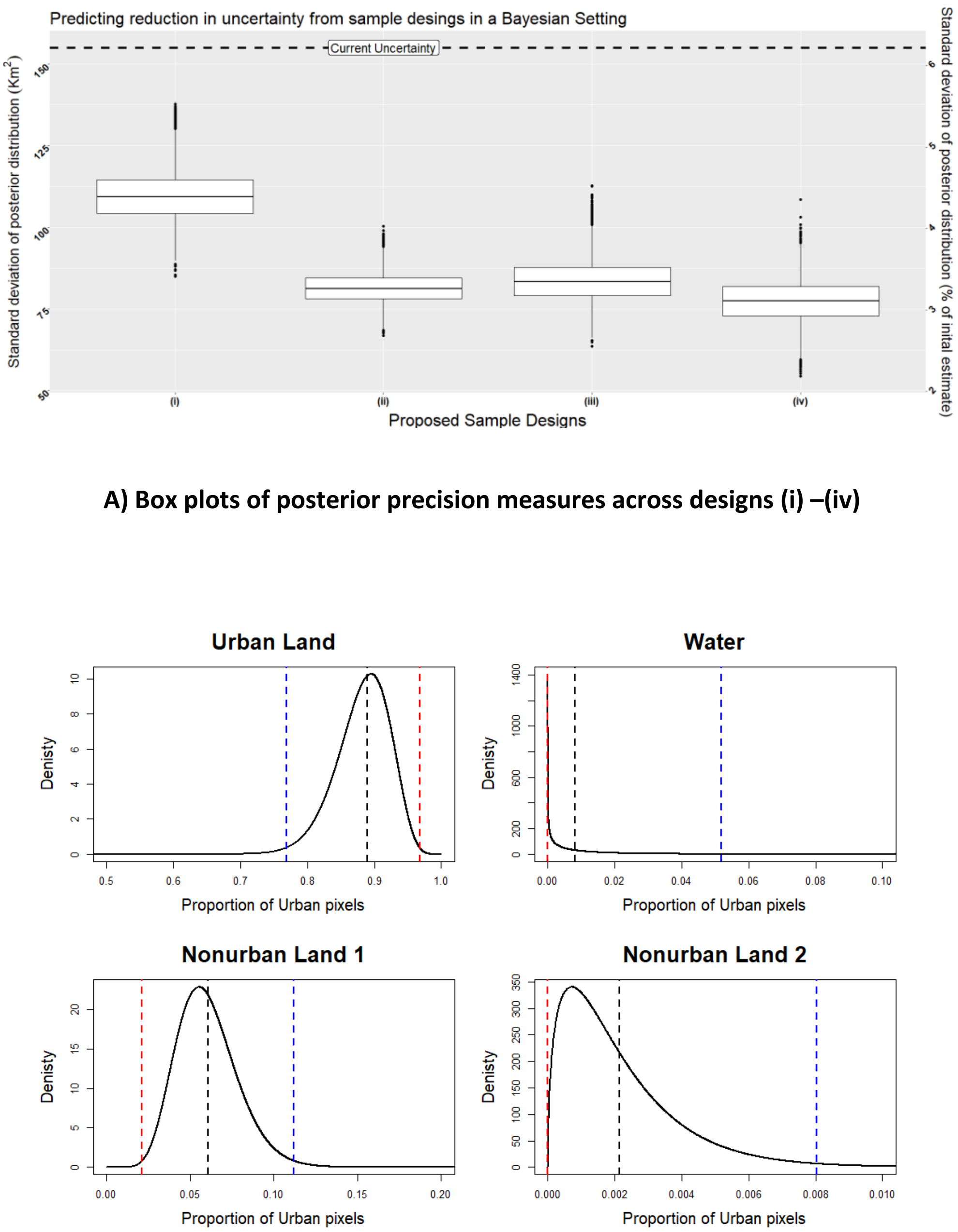
Predicting the effects of further sampling under a Bayesian setting.

When quantifying uncertainty for U under a Bayesian setting, many of the issues seen in the frequentist equivalent can be handled seamlessly. The main reason for this lies in the fact that uncertainty in each p_i are represented as probability distributions, as opposed to intervals which may contain the true value of p_i. This allows one to use methods such as marginalisation and Monte Carlo integration to propagate the uncertainty in each p_i into the posterior distribution for $\sqrt{V\big(A(j)\big)}$. In particular, a sample from the posterior distribution $\sqrt{V\big(A(j)\big)}\ |\boldsymbol{x}, \boldsymbol{x} = (x_1, \dots, x_4)'$may be generated by applying the following procure a large number of times.

1. Set $p_i^*\ i = 1, \dots 4$ where p_i^* is a sample of size 1 drawn from A_i.
2. Generate an artificial sample $\boldsymbol{x}^* = (x_1^*, \dots, x_4^*)'$ where $x_i^* \sim Bin\big(p_i^*, n_{i,j}\ \big)$.

3. Return $\sqrt{V^*} = \sqrt{\sum_{i=1}^{4} \frac{W_i^2\left(x_i+x_i^*+\frac{1}{2}\right)\left(n_i+n_{i,j}-(x_i+x_i^*)+\frac{1}{2}\right)}{\left(n_i+n_{i,j}+1\right)^2\left(n_i+n_{i,j}+2\right)}}$

As a side note, steps 1 and 2 here may also be merged using the beta-binomial distribution [367] in this case to reduce computational demands. The results of this procedure are displayed in Figure 4.6.



A) Box plots of posterior precision measures across designs (i) –(iv)

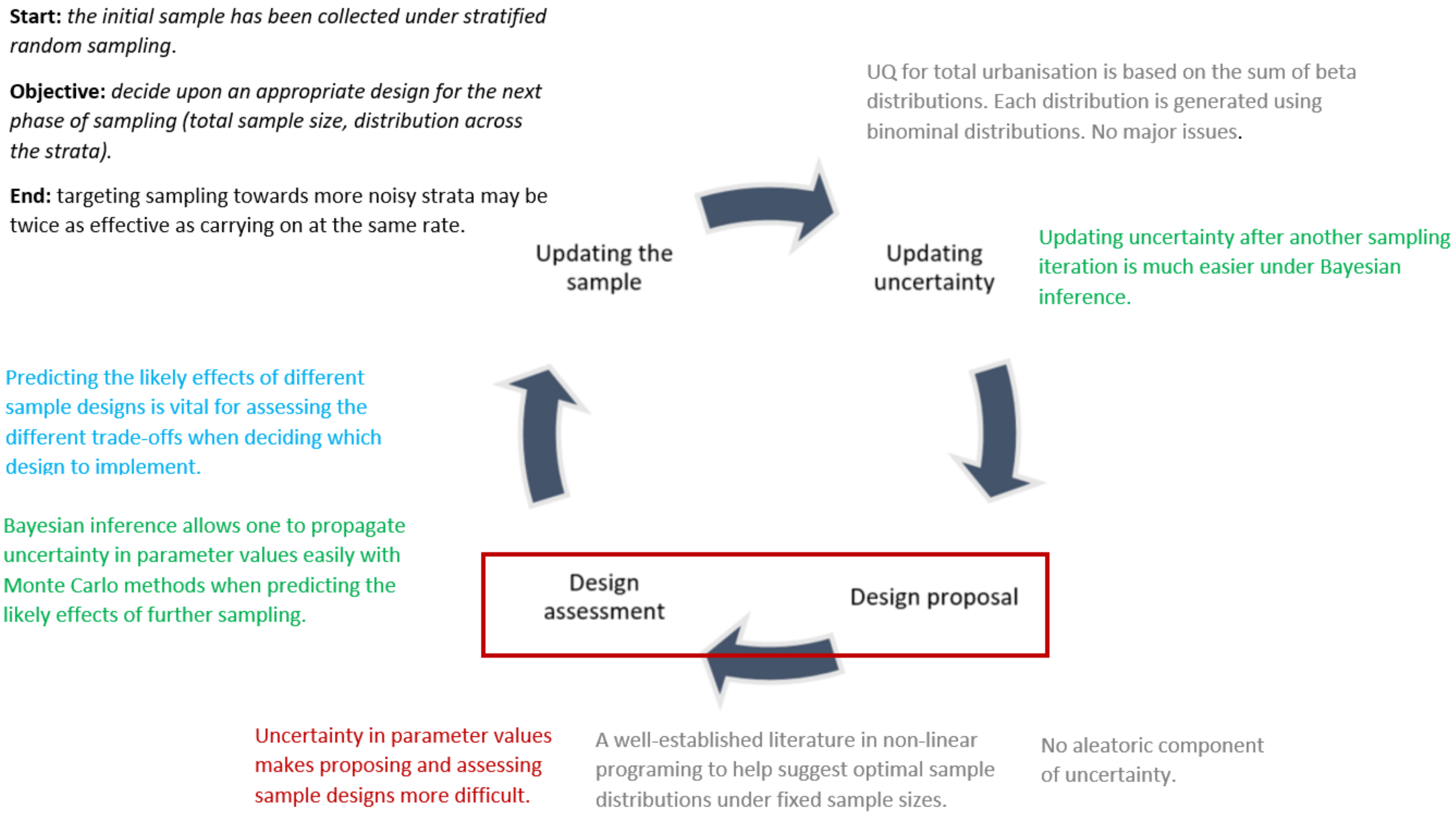
B) Posterior distributions for the proportion of urban pixels in each stratum

Figure 4.6*. (A) Box plots for the precision of the estimated urbanised area based on* 1×10^5 *simulations across the proposed sample designs and posterior distribution in (B). (B) Posterior distributions for the proportion of urban pixels within each stratum from the initial test sample. Coloured lines correspond to the rates assumed in Figure 4.5.*

From the box plots in Figure 4.6A, one can observe that designs (ii) and (iii) are likely to be more efficient than design (i) and may have a similar impact as design (iv) in terms of uncertainty reduction. This result suggests that a targeted sampling of a further 1000 pixels under sample designs such as (ii) or (iii) may be worth just as much as a non-targeted design consisting of a further 3000 pixels when focused solely on urban area estimation. From a performance perspective, it may be difficult to justify (ii) over (iii), as their distributions seem to have a substantial overlap. In a case such as this though, other factors may also be considered. As mentioned earlier, there may be other values that need to be estimated from this sample that may lead a user to favour design (iii). On the other hand, design (ii) avoids any sampling in the nonurban land 2 and water strata, both of which have large areas that are far away from populated areas. Depending on how ground-truths are collected under a more realistic example (e.g. sending experts or using drone footage to collect higher-resolution imagery), avoiding harder-to-reach strata may be a substantial advantage.

4.3 Reflections (Part I)

Overall, the first part of the Lagos case study has highlighted the importance of Bayesian inference when passing through the design assessment phase (see Figure 4.7 for a summary).



Start: *the initial sample has been collected under stratified random sampling.*

Objective: *decide upon an appropriate design for the next phase of sampling (total sample size, distribution across the strata).*

End: targeting sampling towards more noisy strata may be twice as effective as carrying on at the same rate.

Updating the sample

UQ for total urbanisation is based on the sum of beta distributions. Each distribution is generated using binominal distributions. No major issues.

Updating uncertainty

Updating uncertainty after another sampling iteration is much easier under Bayesian inference.

Predicting the likely effects of different sample designs is vital for assessing the different trade-offs when deciding which design to implement.

Bayesian inference allows one to propagate uncertainty in parameter values easily with Monte Carlo methods when predicting the likely effects of further sampling.

Design assessment

Design proposal

Uncertainty in parameter values makes proposing and assessing sample designs more difficult.

A well-established literature in non-linear programing to help suggest optimal sample distributions under fixed sample sizes.

No aleatoric component of uncertainty.

Figure 4.7. *A summary of how the adaptive sampling framework has helped to overcome the challenges in the first part of the Lagos case study presented in Figure 4.1*

In particular, the first part of this case study has illustrated that predicting the likely effects of different sample designs is a lot more interpretable and generalisable under Bayesian inference when compared to frequentist inference. The additional interpretability comes from the fact that Bayesian inference allowed one to fold uncertainty in other parameter values into the likely effects of further sampling to give a single distribution of each design which made comparing the likely impact of different sample designs much easier.

The additional generalisability is a result of the ability to combine marginalisation and Monte Carlo integration. In this part of the case study, one was first able to use marginalisation to break the problem of predicting the likely effects of further sampling into more manageable sub-tasks. These sub-tasks involved (i) quantifying the uncertainty in parameter values (this was done by quantifying uncertainty in each parameter p_i using A_i) (ii) replicating each sample design under fixed parameter values (this was done using a series of binomial distributions) and (iii) updating the chosen precision measure once an additional set of reference data had been sampled (this was done by using (4.2)). These three subtasks can then be combined to predict the likely effects of different sample designs via Monte Carlo integration.

Apart from one minor step which allowed the first two sub-tasks to be merged with beta-binomial distributions to improve computational efficiency, there was nothing specific about the model or sample designs that was vital in these three steps. Hence, this combination of marginalisation and Monte Carlo integration may be used to predict the likely effects for a wide variety of models and sample designs providing that one can (i) quantify the uncertainty in parameter values, (ii) replicate sample designs under fixed parameter values and (iii) update uncertainty in an estimate under a set of data collected under a sample design.

Looking towards the adaptive framework, combinations of marginalisation and Monte Carlo integration play very important parts in ensuring the generalisability of the framework as a whole. Given that these simulation-based methods rely on Bayesian inference, the use of Bayesian inference becomes almost mandatory in an adaptive sampling framework.

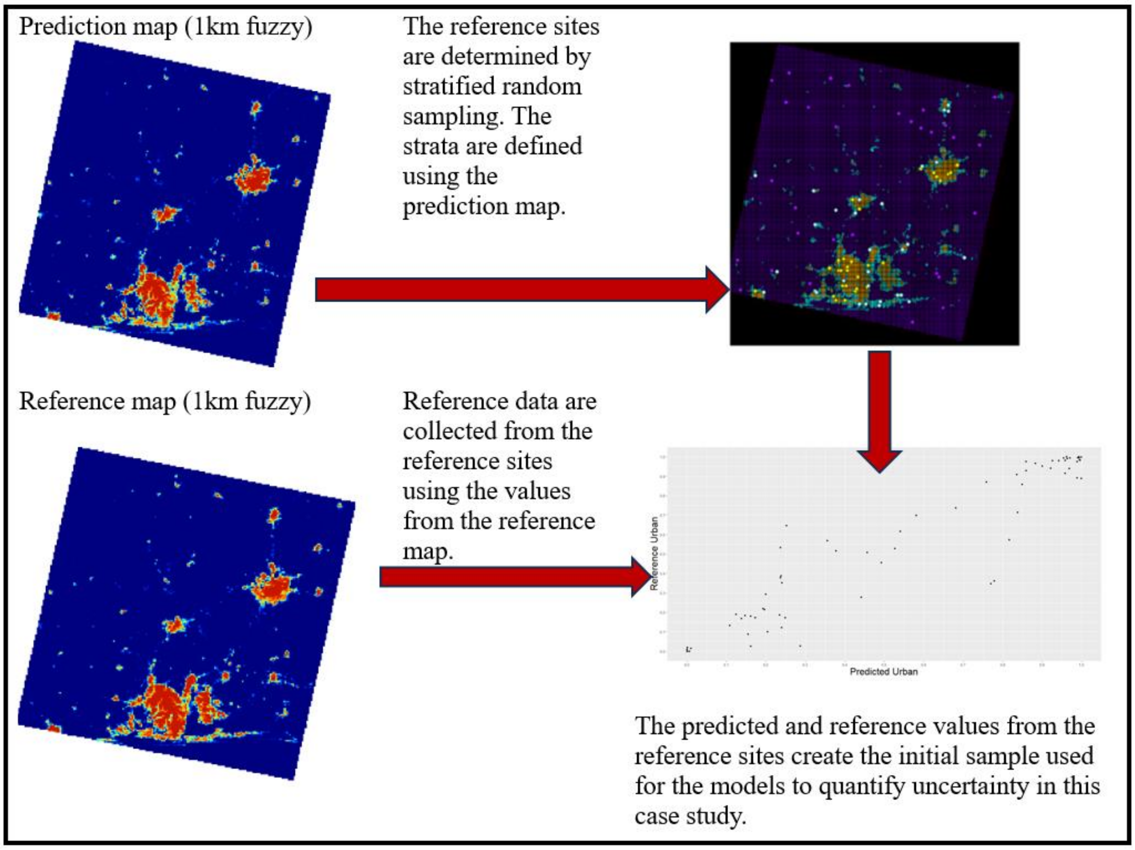
From a more general perspective, it is possible to abstract from the first part of the case study to UQ for estimates made from confusion matrices. Some examples here include many commonly used performance metrics such as estimating sensitivity, specificity, and overall accuracy. Furthermore, the methods used in this part of the case study were agnostic to how the map was generated, meaning one would have been free to use any other classification techniques (machine learning or otherwise).

4.4 Lagos urban mapping (II): quantifying uncertainty for individual cases in a fuzzy classification

Part II of the Lagos case study evaluates the framework under a fuzzy classification problem that quantifies uncertainty for individual instances (in this case 1km pixels). This differs from part I, which focused on quantifying uncertainty for a global level estimate (i.e. the total urbanised area) based on discrete classifiers.

Here, each 1km square pixel is assigned a predicted and reference urbanisation extent score based on the proportion of pixels classified as urban in the 30m resolution map that falls with each 1km pixel. For example, if 25% of the 30m pixels in a 1km area on the prediction map were classified as Urban then the predicted extent would be 0.25.

Under this part of the case study, an initial sample size of 90 is collected under stratified random sampling based on the predicted extent of urbanisation. Specifically, 30 pixels are randomly selected from each of the following strata: low urbanised areas (a predicted extent of urbanisation below 0.1), semi-urbanised areas (a predicted extent of urbanisation between 0.1 and 0.7) and highly urbanised areas (a predicted urban extent greater than 0.7). This initial sample design was chosen as 30 pixels from each stratum was deemed a suitable amount of reference data for an initial sample to be informative without being excessive.



The reference sites are determined by stratified random sampling. The strata are defined using the prediction map.

Reference data are collected from the reference sites using the values from the reference map.

The predicted and reference values from the reference sites create the initial sample used for the models to quantify uncertainty in this case study.

Figure 4.8 *An overview of how the initial sample is obtained in the second part of the Lagos case study.*

From this initial sample, a truncated Gaussian model is fitted between the predicted and reference urban extents with

$$y_i \sim N_T\left(\beta_0 + \beta_1 x_i, \sigma^2\left(1 + 4\alpha\big(x_i(1 - x_i)\big)\right)^2, \min = 0, \max = 1\right), \tag{4.4}$$

where y_i, x_i are the ground-truth and predicted values for the urbanised area pixel i respectively and $N_T(\mu, \sigma^2, \min = a\,, \max = b)$ denotes a truncated normal distribution based on a normal distribution of mean μ and variance σ^2 that is bounded within the interval (a, b).

Posterior distributions are generated using the Metropolis-Hastings algorithm [368] with the likelihood function derived from the truncated normal distribution [369] and leveraging the independence of errors. Here, a non-informative prior is placed on $\boldsymbol{\theta} = (\beta_0, \beta_1, \sigma^2, \alpha)'$ with $\pi(\boldsymbol{\theta}) \propto \frac{1}{\sigma^2}$.

Effectively, this model is an altered form of a standard linear regression model that has been adapted to consider two additional factors. Firstly, the switch to a truncated normal distribution allows one to factor in the fact that the reference data are bounded between 0 and 1. Secondly, a scaling factor of $1 + 4\alpha\big(x(1-x)\big)$ has been placed on the variance to allow for heterogeneous errors that deflate when the predicted values are close to 0 or 1 and to inflate around 0.5, with α controlling the degree of this inflation. This allows one to factor in that disagreement between predicted values and reference values are expected to be smaller around homogeneous areas and greater in heterogeneous areas.

Figure 4.9 provides a difference plot between the reference and predicted values for the extent of urbanisation along with a comparison of the standard linear regression model and the truncated normal model proposed in (4.4) fitted to the reference sample. Here, Figure 4.9 suggests that a standard linear model (which assumes constant variance) is too simplistic whereas a model truncated normal model (4.4) is more appropriate.

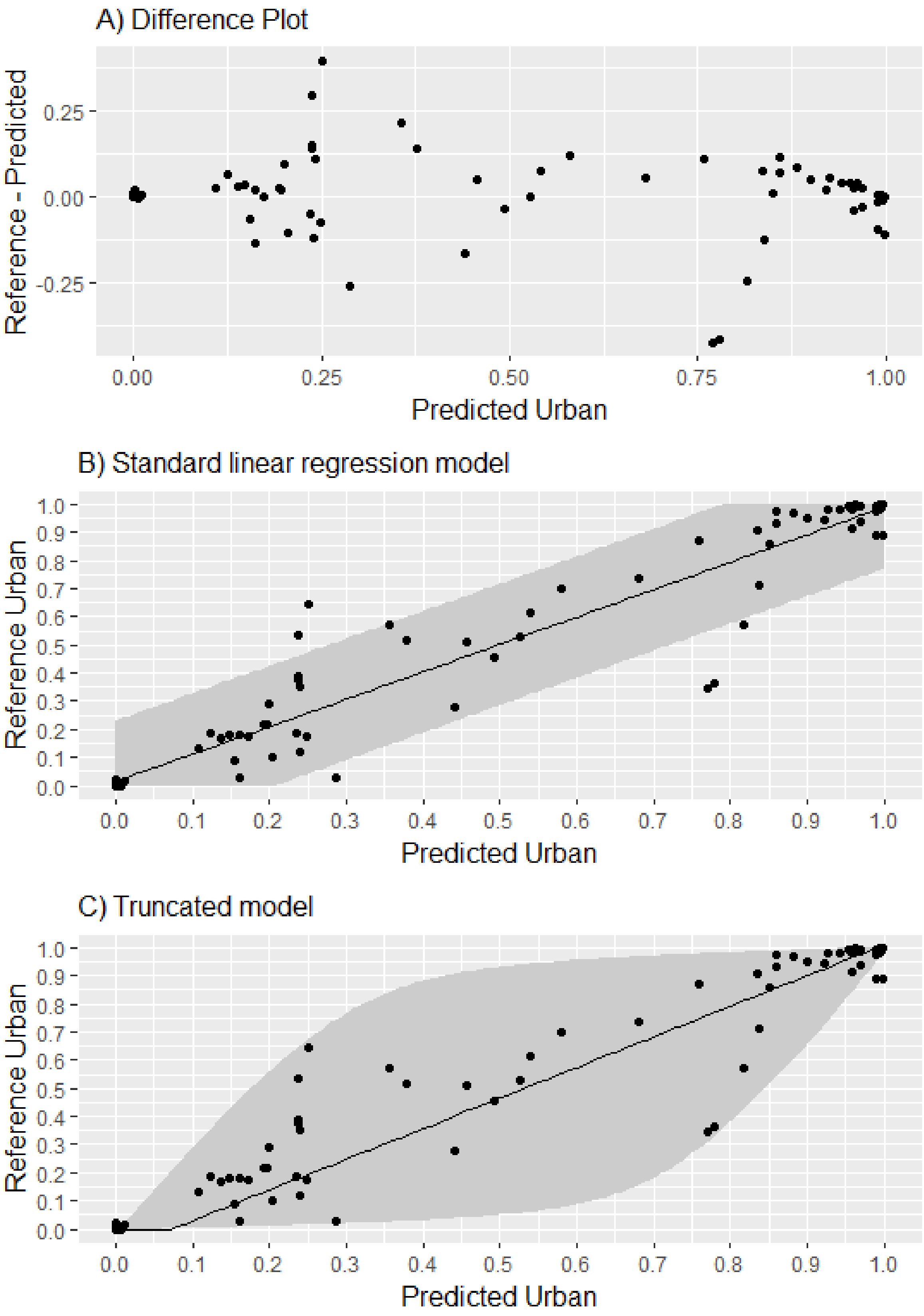


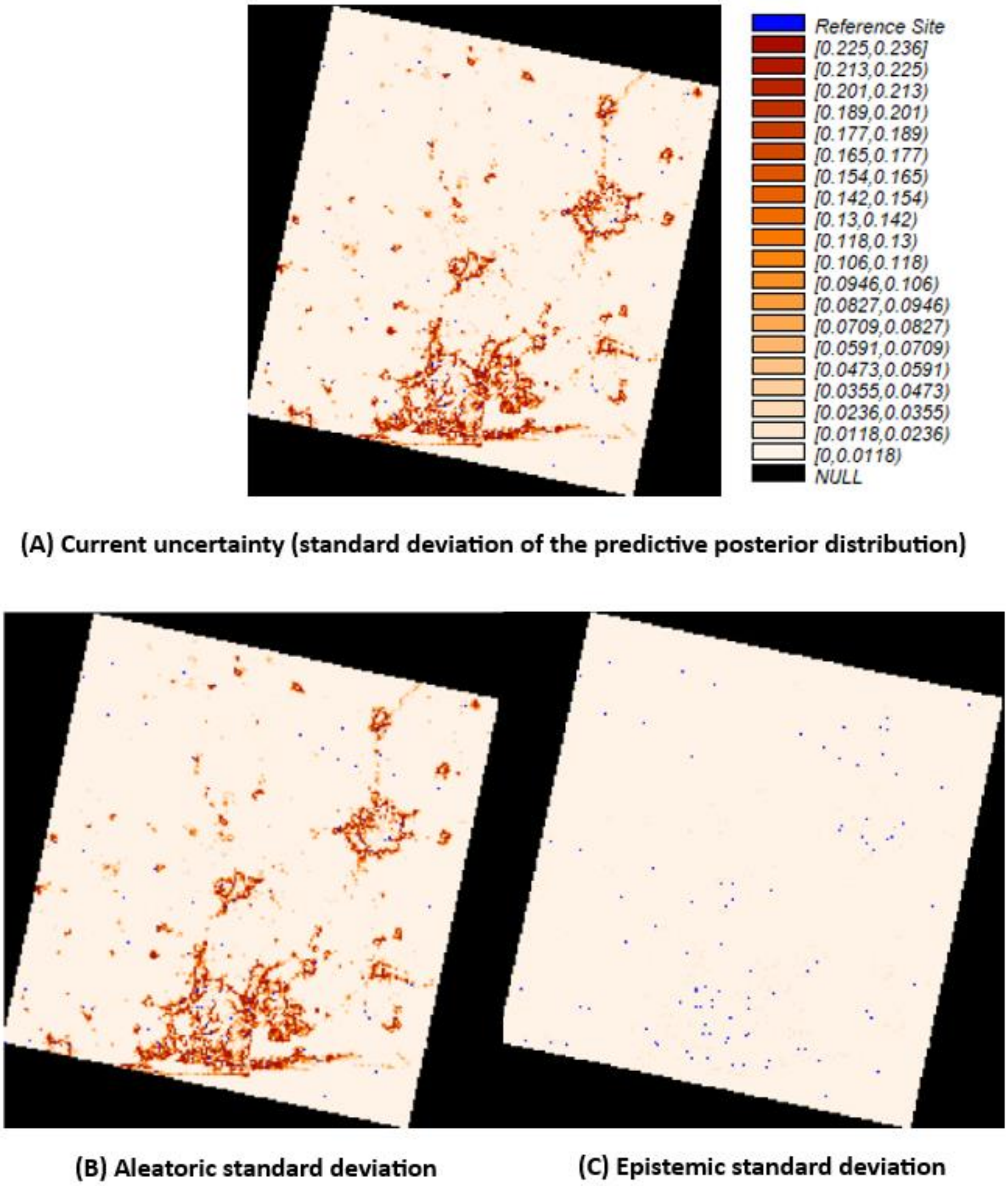
Figure 4.9 *(A) A difference plot between the reference and predicted urbanisation as the predicted value varies. (B) A simple linear regression fitted to the observed data set. (C) A truncated and inflated model proposed in (4.4). The grey bands in B and C represent equal-tailed 95% prediction intervals for each point under each model.*

Proposing efficient sample designs under the model in (4.4) is more complex when compared to the area estimation problem in the first part of the case study, as there is now an aleatoric component to uncertainty governed by unknown parameters σ and α along with the predicted urban extent for each 1km pixel. Hence, there is now an upper bound as to how far uncertainty may be reduced by further sampling, with this bound itself subject to a degree of uncertainty varying across the map.

Fortunately, the issue of uncertainty in the aleatoric (and epistemic) components of uncertainty may be dealt with relatively easily when UQ is based on Bayesian inference through marginalisation and Monte Carlo integration.

Figure 4.10 displays estimates for the aleatoric and epistemic standard deviations across the map, whilst Figure 4.11 compares the current and aleatoric uncertainty across the predicted urban extents.

Both plots suggest that further sampling is likely to have little impact on reducing uncertainty in this model, as the level of uncertainty and the aleatoric component of uncertainty seem to be close (or equivalently, the epistemic component seems close to 0).



(A) Current uncertainty (standard deviation of the predictive posterior distribution)

(B) Aleatoric standard deviation

(C) Epistemic standard deviation

Figure 4.10 *(A) standard deviations for predictive posterior distribution for each 1km pixel (i.e. the current level of uncertainty). (B) and (C) map the mode of aleatoric standard and epistemic standard deviations respectively.*

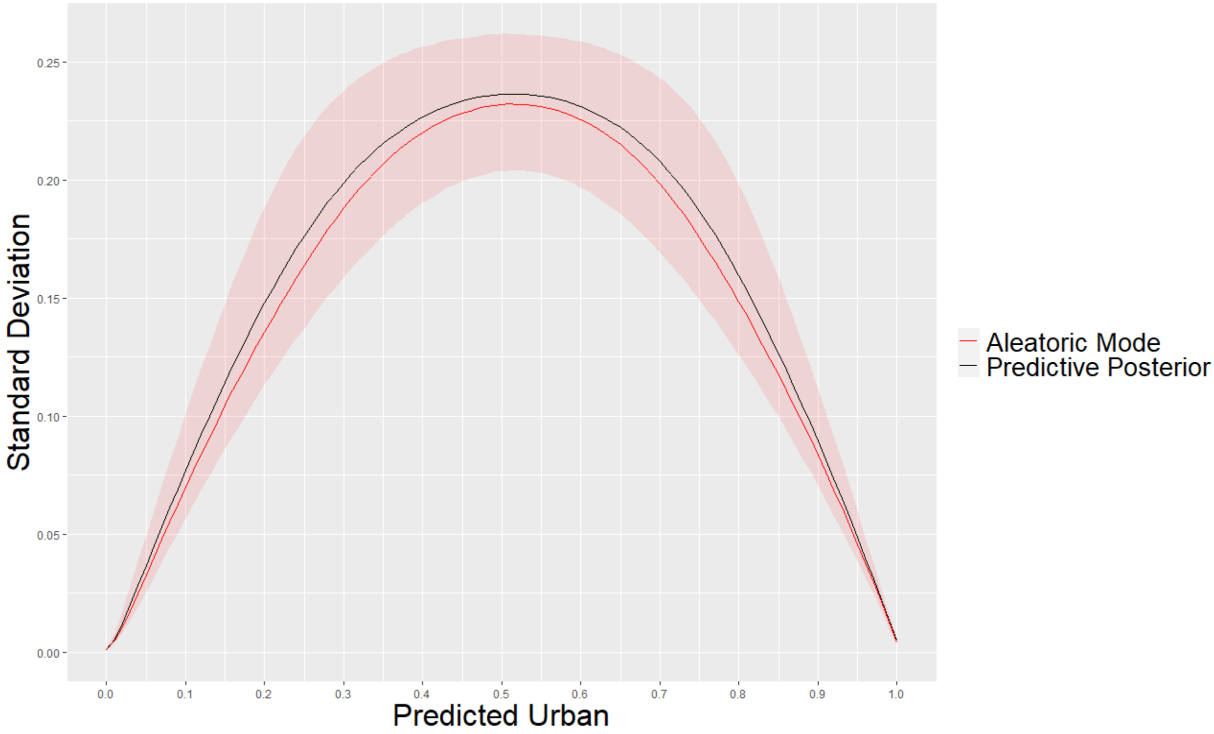


Figure 4.11*. The predictive standard deviations (black) along with the mode of aleatoric standard deviation as the predicted degree of urbanisation in the 1km squares vary. The red band represents an equal-tailed 95% credible interval for the aleatoric standard deviation at each point.*

As side notes, many alternative model structures to (4.4) also account for bounded and heterogeneous errors. For example, one alternative way to account for bounded values would be to use a two-sided Tobit model [315]. Equally, several modelling assumptions in (4.4) could have been validated further (e.g. the independence of error terms, the suitability of a Gaussian distribution etc.).

In practice, one may wish to consider a range of alternative models and validate assumptions using formal statistical testing. For the purpose of this thesis though, this case study does not go to this level of detail in assessing model choice and the validity of assumptions.

The reason for this is that the focus for this part of the case study is on using Monte Carlo methods to quantify aleatoric and epistemic components of uncertainty once a model has been chosen. Including multiple alternative models is likely to only repeat many of the steps in the evaluation. Likewise, validating every modelling assumption places a considerable amount of work for a thesis of this nature on issues that are not its primary focus. In other words, this thesis moves past the model choice and assumption validation steps fairly quickly in order to focus on the steps after a model has been agreed upon.

As a secondary side note, the variance of a truncated normal distribution has a closed-form solution, which is given by

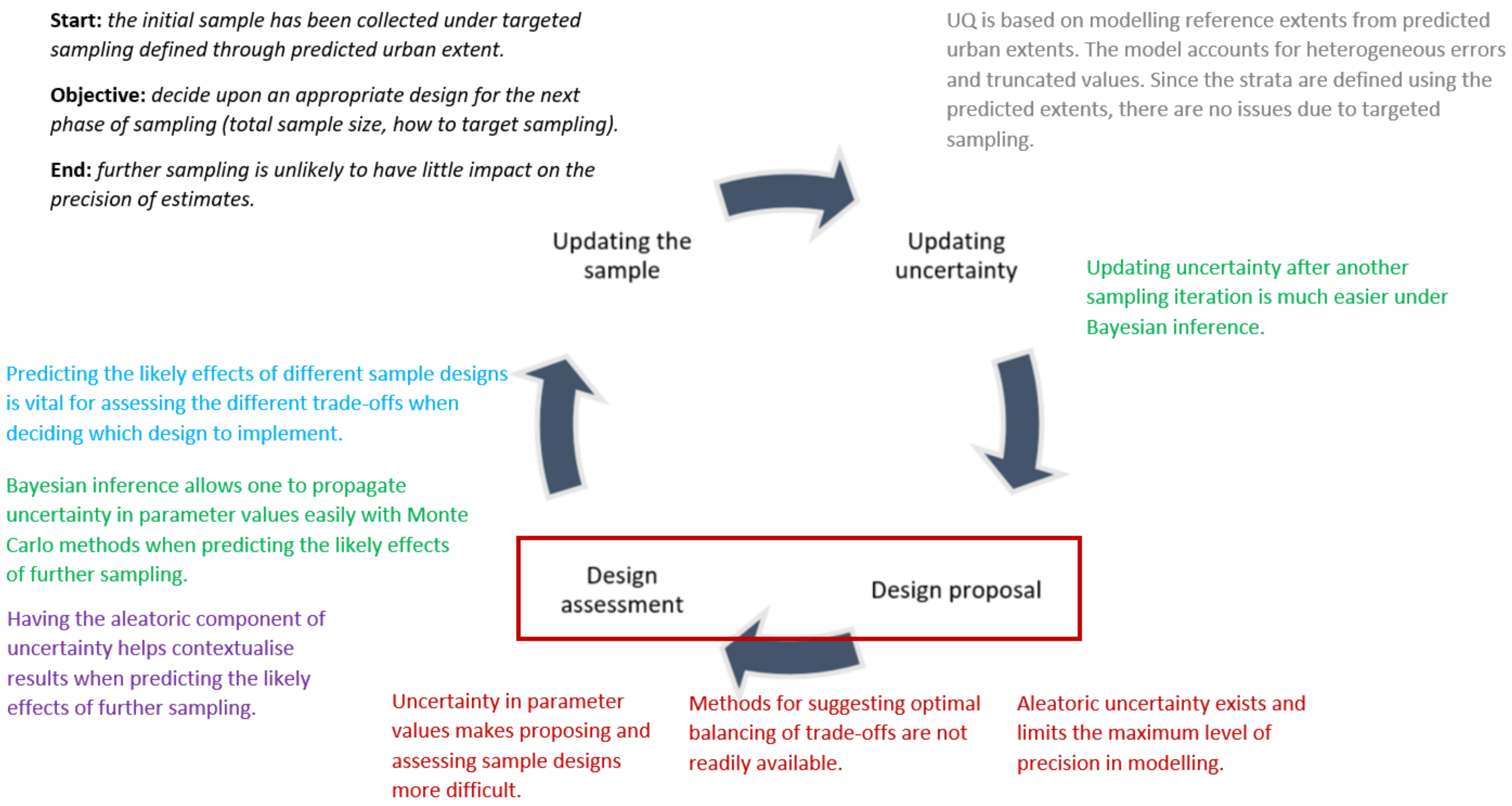
$$X \sim N_T(\mu, \sigma^2, \min = a\,, \max = b) \;\Rightarrow\; V(X) = \sigma^2 \left[1 + \frac{A\phi(A) - B\phi(B)}{Z} - \left(\frac{\phi(A) - \phi(B)}{Z}\right)^2\right],$$

where $A = \frac{a-\mu}{\sigma}, B = \frac{b-\mu}{\sigma}\ Z = \Phi(B) - \Phi(A)$; ϕ, Φ denote the probability and cumulative density functions for a standard normal distribution $N(0,1)$ respectively.

Closed-form solutions such as this help in making marginalisation and Monte Carlo easier by reducing computational demands. Strictly speaking though, a closed-form solution is not necessary when quantifying aleatoric and epistemic variances (or standard deviations) as variances themselves can be estimated through Monte Carlo integration under fixed parameter values.

4.5 Reflections (Part II)

The second part of the Lagos case study has illustrated how estimating aleatoric and epistemic components of uncertainty can help users determine when one is approaching the limits of further sampling under a fixed model structure (see Figure 4.12 for a summary).



Start: *the initial sample has been collected under targeted sampling defined through predicted urban extent.*

Objective: *decide upon an appropriate design for the next phase of sampling (total sample size, how to target sampling).*

End: *further sampling is unlikely to have little impact on the precision of estimates.*

UQ is based on modelling reference extents from predicted urban extents. The model accounts for heterogeneous errors and truncated values. Since the strata are defined using the predicted extents, there are no issues due to targeted sampling.

Updating the sample

Updating uncertainty

Updating uncertainty after another sampling iteration is much easier under Bayesian inference.

Predicting the likely effects of different sample designs is vital for assessing the different trade-offs when deciding which design to implement.

Bayesian inference allows one to propagate uncertainty in parameter values easily with Monte Carlo methods when predicting the likely effects of further sampling.

Having the aleatoric component of uncertainty helps contextualise results when predicting the likely effects of further sampling.

Design assessment

Design proposal

Uncertainty in parameter values makes proposing and assessing sample designs more difficult.

Methods for suggesting optimal balancing of trade-offs are not readily available.

Aleatoric uncertainty exists and limits the maximum level of precision in modelling.

Figure 4.12. *A summary of how the adaptive sampling framework has helped to overcome the challenges in the second part of the Lagos case study presented in Figure 4.2.*

Like part I, this second part of the Lagos case study relied heavily on being able to apply marginalisation and Monte Carlo integration (which was only possible because one adopted Bayesian inference in UQ). This time though, marginalisation and Monte Carlo integration were needed to estimate components of uncertainty when no closed-form solutions were available.

From a more general perspective, the ability to estimate the limits of further sampling plays a vital part when managing uncertainty, as it allows one to contextualise any results at the design assessment phase and avoid committing resources to inefficient sample designs. Effectively, being able to estimate the aleatoric component of uncertainty gives users a way to know when it is time to stop sampling (or when to stop targeting some areas of a population).

With both parts of this case study completed, one can begin to consider how the lessons learned under this Lagos example may be brought forward into the second case study. In particular, three observations carry forward to the next case study.

Bayesian inference is vital for ensuring generalisability

In both parts of the Lagos study, one relied upon a combination of marginalisation and Monte Carlo integration to propagate uncertainty in key predictions and estimations. In the area estimation problem from the first part of the case study, this method was used to predict the likely effects of different sample designs. In part II, this method was used to estimate the aleatoric component of uncertainty for each 1km pixel. Ultimately, such combinations of marginalisation and Monte Carlo are only possible under Bayesian inference. Hence whilst it may be possible to predict the likely effects of sampling and estimate components of uncertainty in some frequentist settings, choosing to do so may come at the cost of foregoing highly generalisable ways of applying such methods.

Sample design tends to involve balancing probabilistic trade-offs rather than optimising designs

Another trait common to both parts of this Lagos case study is that it was impossible to give an objective answer as to which sample designs were more appropriate. Even in the relatively simple case study part I (where one had access to nonlinear programming to suggest optimal designs), the best one was able to do was to suggest the likely impacts of different sample designs (see Figure 4.7). Given the overlap between the distributions due to uncertainty in the parameters and stochastic variation, along with considerations outside estimating the urban areas (e.g. reducing uncertainty for other metrics, spatial clustering etc.), there was a degree of subjectivity as to what the best sample design would be in this case.

In the second part of the case study, the additional model complexity means that it was difficult to formulate efficient sample designs in the same way that it was possible in the area estimation problem in part I. However, one was still able to use the estimates for the aleatoric component of uncertainty to conclude further sampling was unlikely to do much in the way of reducing uncertainty in the urban extent map. The key thing to note here is that one could express what would be likely to happen given the current data. This is illustrated in Figure 4.11 where the aleatoric component is given as a band.

Ultimately, these two parts of the case study highlight that managing the relationship between uncertainty and reference sampling is itself subject to uncertainty and typically involves balancing trade-offs with probabilistic statements as opposed to optimising sample designs based on strict criteria.

There is an unexplored method of generating sample designs via clusters of epistemic uncertainty in the predictive feature space

Under part II of this case study, one was able to view how the aleatoric (and epistemic) components of uncertainty varied across the map and predicted urban extent. The motivation behind this is that clusters of high epistemic uncertainty may be useful for informing targeted sampling practices to manage uncertainty efficiently. In this example though, there were no such clusters, as most of the uncertainty after the initial sample was aleatoric across each pixel in the map.

Nevertheless, the question is still raised as to what one would have done if clusters of high epistemic uncertainty were present. For example, if one were to notice pixels with a high epistemic component cluster spatially across the map, the solution may not be as simple as simply targeting these areas when determining the design for the next iteration. From Section 3.2.3 though, one knows UQ is made a lot easier if the reference data are obtained under a probabilistic design defined through the predictors of the model used to quantify uncertainty.

This motivates a hypothesis that one good way of generating efficient sample designs would be to look for clusters of high epistemic uncertainty in the predictive feature space and to target any probabilistic sampling more heavily toward these areas.

Unfortunately, it was difficult to test this hypothesis in part II of the Lagos mapping problem, as such clusters did not exist after the first iteration of sampling. Nevertheless, this idea is something one would want to bring forward into the second case study.

When combining these three observations and looking toward the second case study, this thesis proposes that a large part of any solutions in the second case study will involve trying to find clusters of high epistemic uncertainty and looking to target them through probabilistic sampling within the predictive feature space. However, one suspects that challenges such as uncertainty in key parameter values and other considerations (e.g. a preference for spatial clustering) will make the problem of efficient sampling less than straightforward. Instead, one suspects that plots like Figure 4.9 and Figure 4.10 (which compare components of uncertainty across spatial domains and predictive features) will play a large role in developing cost-effective sample designs and will be used alongside analysis rooted in Bayesian inference that will express results using probabilistic statements.

4.6 Summary

This chapter has considered the Lagos case study as the first stage of evaluating the adaptive sampling framework and has been split into two parts. The first part of the case study considered the problem of optimising sample designs when estimating the total urbanised area based on a stratified random sampling of discreetly classified pixels. The second part of the case study considered how to best manage the cost-benefit trade-off in reducing uncertainty for individual pixels under fuzzy classification.

For the first part of the case study, the fundamental issue was that one needed to know how to distribute sample sizes across the strata when estimating the total urbanised area. Here, there were two factors to consider: the relative size of each stratum and the true proportion of urban pixels in each stratum. With the proportion of urban pixels in each stratum assumed to be unknown (although in this case, one did have a full reference map), there was an additional layer of uncertainty as to how different sample designs may impact the precision of area estimates.

Ultimately, the switch to UQ under Bayesian inference allowed one to incorporate these sources of uncertainty. From this, one was able to provide probabilistic statements for the likely impacts of sample designs before any further sampling had taken place. Effectively, this allowed a *try-before-you-buy* approach to sample design, which may be used to provide users with the assurances they may need when comparing different sample design options under uncertainty.

For the second part of the case study, uncertainty for individual pixels at the 1km level was quantified by fitting a model between their predicted and reference values. As is the case with many models of this nature, there was an aleatoric component of uncertainty, which needed to be estimated to ensure one would not waste resources committing to sample designs that were likely to have little or no effect on reducing uncertainty. Once again, the true nature of the aleatoric component was subject to uncertainty because of unknown parameter values, which were incorporated into the analysis via Bayesian inference. The ability to estimate the aleatoric component of uncertainty helped greatly in this part of the case study as it gave one a means of estimating the maximum level of precision for different pixels under the model, which allowed one to conclude that enough reference data had already been sampled (under the chosen model).

When reviewing both parts of the case study, three themes became clear. Firstly, the decision to use Bayesian inference in UQ greatly increases the generalisability of techniques used in adaptive sampling, as it allows one to take advantage of simulation-based methods when closed-form solutions are difficult to obtain. Secondly, adaptive sampling is better viewed as a way to efficiently manage trade-offs between sampling costs and uncertainty reduction rather than trying to treat it as a way of solving optimisation problems. Thirdly, the ability to distinguish between aleatoric and epistemic uncertainty is likely to have an important role in managing these trade-offs, particularly in more model-dependent forms of UQ.

Overall, the Lagos case study has acted as a useful first test for evaluating the proposed framework and methods. Broadly speaking, the framework has been a success in this context. However, in some sense, the Lagos case study represented an easier level of challenge by having a full reference map available from the outset. The next phase of the analysis (i.e. Chapter 5) will test the framework on a case study with challenges related to sampling bias and propensity scoring.

Chapter 5 Case Study 2: England Woodland Mapping

5.1 Problem introduction.

For the most part, the proposed adaptive framework and methods performed well under both parts of the Lagos case study. The purpose of this chapter is to evaluate the adaptive framework under a more difficult set of circumstances involving sample bias and where there is a preference to sample from some areas more than others due to sample restrictions (i.e. propensity scoring).

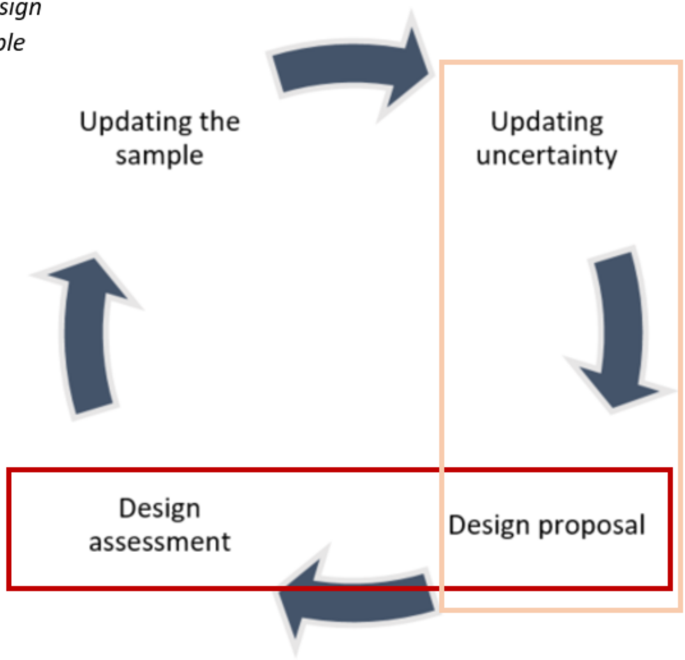
More specifically, Chapter 5 evaluates the proposed adaptive sampling framework on the England woodland mapping example first introduced in Chapter 1. This England study differs from the Lagos study in three major ways. Firstly, there is no longer the luxury of a full reference map. This is something one would expect to see in most real-world applications. Secondly, the England case study begins with the initial set of reference data sampled under a known bias. Thirdly, there is a general need to avoid some areas of the map and favour others when sampling, which is captured using propensity scores.

This case study has been chosen to represent two problems common to adaptive sampling: (i) initial samples may be biased due to practical restrictions and (ii) sampling costs can often vary across a population which needs to be considered when balancing trade-offs between sample designs and uncertainty reduction.

From the perspective of the adaptive sampling framework, the problem of initial sample bias and varying propensity scores creates additional challenges in the updating uncertainty and design proposal stages (see Figure 5.1 for further details). Under this perspective, the purpose of this chapter is to provide a case study to evaluate how the methods set out in Chapter 3 fare under these additional challenges.

Start: *the initial sample has been collected under targeted sampling defined through a propensity score.*

Objective: *decide upon an appropriate design for the next phase of sampling (total sample size, how to target sampling).*



Updating the sample

Updating uncertainty

Design proposal

Design assessment

The propensity score determines bias in the initial sample. This may cause issues in UQ if not accounted for.

Design restrictions related to propensity scoring makes proposing efficient sample designs more difficult.

Uncertainty in parameter values makes proposing and assessing sample designs more difficult.

Methods for suggesting optimal balancing of trade-offs are not readily available.

Aleatoric uncertainty exists and limits the maximum level of precision in modelling.

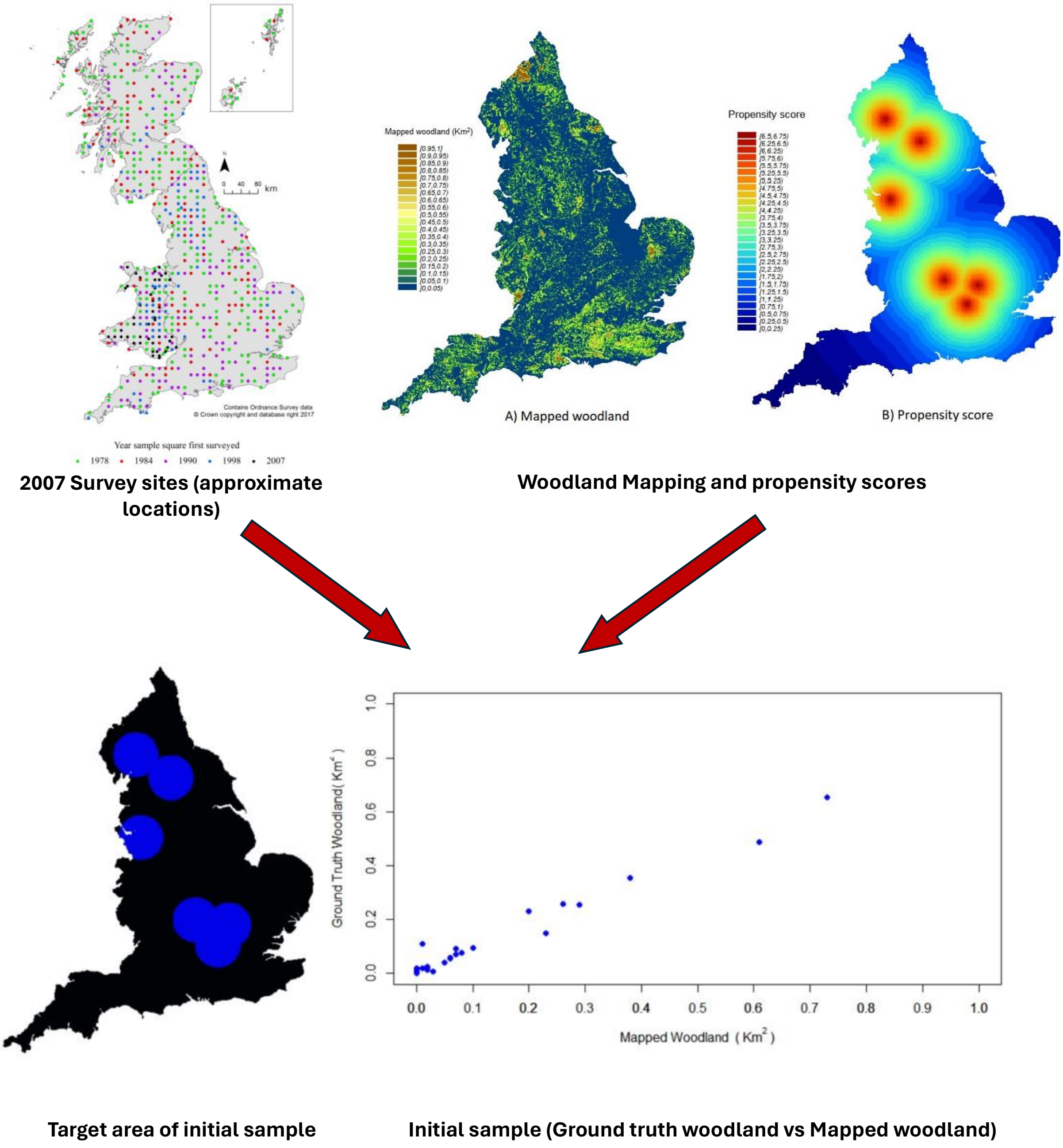
Figure 5.1. *A summary of how the England woodland mapping study relates to challenges across the four key stages of adaptive sampling.*

For this woodland case study, the prediction map is constructed by first dividing the area of England into approximately 130,000 1km pixels. Following this, each pixel is assigned a predicted proportion of woodland area based on the proportion of the 25m pixels in the area that are classified either as *Broadleaved, Mixed and Yew Woodland,* or *Coniferous Woodland* according to the 2015 UK land cover map [371].

As stated earlier, one major difference between this second case study and the Lagos case study is in the reference data. Here, there is no full reference map. Instead, the reference values (or ground-truths) are based on real-world surveys. The ground-truth values for the total proportion of woodland area are then extracted from these 21 class surveys.

For this case study, the initial sample is based on a truncated set of the original reference data from the 2007 Countryside Survey data [372], [373]. When the 2007 survey data were collected, systematic random sampling was applied based on 15km spatial grids that covered the entirety of England. However, this survey was conducted when mappings based on machine learning and satellite imagery were arguably still in the early stages. As the techniques in machine learning and the availability of satellite imagery have developed, it has become increasingly cheaper to produce prediction maps at higher temporal resolutions. This, in turn, has put much more attention on the costs and timeliness of collecting survey data. In short, relying on large countrywide surveys every decade or so for UQ in maps creates a bottleneck in the overall mapping process that becomes increasingly relevant as it becomes easier to produce the prediction maps cheaply and at higher time resolutions. The need to move away from large countrywide surveys has been further motivated in recent years as a result of travel restrictions due to COVID-19 regulations.

At the time of this thesis, no real-world survey data sampled under COVID-19 travel restrictions currently exists. To get around this, this case study used the historical 2007 survey data and supposed a hypothetical set of travel restrictions based on distances from surveyors' homes. To put this another way, this case study considers the problem of how one would have selected survey sites in 2007 if faced with travel restrictions similar to the COVID-19 regulations. Under this hypothetical example, the initial sample is based on a small sample of (n=30) close to where the surveyors are based (a propensity score greater than or equal to 4). This process is summarised in Figure 5.2. In this case study, the role of the adaptive sampling framework is to help users better manage the cost-benefit trade-offs that may come with travelling to less desirable sites to conduct surveys.



2007 Survey sites (approximate locations)

Woodland Mapping and propensity scores

Target area of initial sample

Initial sample (Ground truth woodland vs Mapped woodland)

Figure 5.2 *An overview of the initial sample and its design for the England case study. Here the reference data are based on a 2007 survey. The initial sample is a random selection of 30 survey sites that fall within the initial target area (i.e. The blue areas on the bottom left map, which represent the areas with propensity scores greater than or equal to 4).*

5.2 Updating the uncertainty from a biased sample

With the initial sample already collected, the first task in the adaptive sampling framework in this case study is to quantify uncertainty for the first time. Unfortunately, fitting a model between the predicted and reference values without accounting for the bias is not going to be possible without some heavy modelling assumptions (see Section 2.2.1 for further details). In this instance, the assumptions would require that the predicted woodland extent somehow causes the ground-truth values to change. Clearly, this is not the case (as much as a world where increasing woodland habitats was as easy as editing the predicted values in a map does seem appealing).

Consequently, one must be able to account for bias in the sample design when using a model to quantify uncertainty. Given one knows that the bias in the initial sample is dependent on only the propensity score, one can use the results in Section 3.2.3 to account for bias in the design by incorporating the propensity score into the model.

For the England woodland study, a data-driven model for the reference woodland values (given the predicted values from the map) is based on a Bayesian kernel machine regression model. Here, the model for reference woodland value of pixel y_i given its predicted value x_i and propensity score c_i is written as

$$y_i = \beta_0 + \beta_1 x_i + h(\boldsymbol{z}_i) + \epsilon_i \tag{5.1}$$

where $\epsilon_i \sim N(0, \sigma^2)$, $\boldsymbol{z}_i = (x_i, c_i)'$, $h(\cdot)$ is some flexible function based on kernel machine regression under a Gaussian kernel function (see [374] for further details).

The motivation for this model is based on two principles. The first principle is that one suspects the predicted and reference woodland extents to be positively correlated. This motivates the linear component, $\beta_0 + \beta_1 x_i$. The second principle is that one would like a reasonable degree of flexibility in this data-driven model to avoid too many issues over ontological uncertainty and ensure there is enough epistemic uncertainty to make adaptive sampling worthwhile. This motivates the second kernel-based component, $h(\boldsymbol{z}_i)$.

Once again, there are many alternative models with flexible structures one could have used (e.g. generalised additive models [203], Gaussian process models [375] etc.) and even models that consider spatial autocorrelation structures [376]. For reasons similar to those discussed at the end of Section 4.4, a full assessment of modelling choices is omitted at this stage in order to swiftly move on to the next stages in adaptive sampling. In general, though, it is good practice to assess the sensitivity of model choice and thoroughly validate modelling assumptions.

Under the initial sample, the model in (5.1) is fitted. Figure 5.3 displays this model fit along with a 95% prediction surfaces view across the propensity scores and predicted woodland extents.

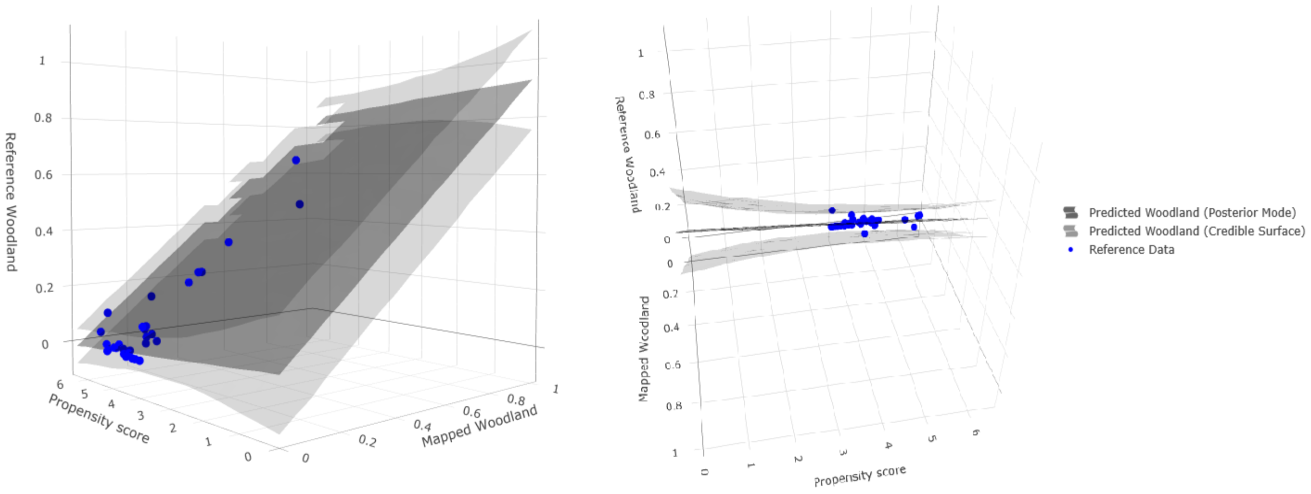
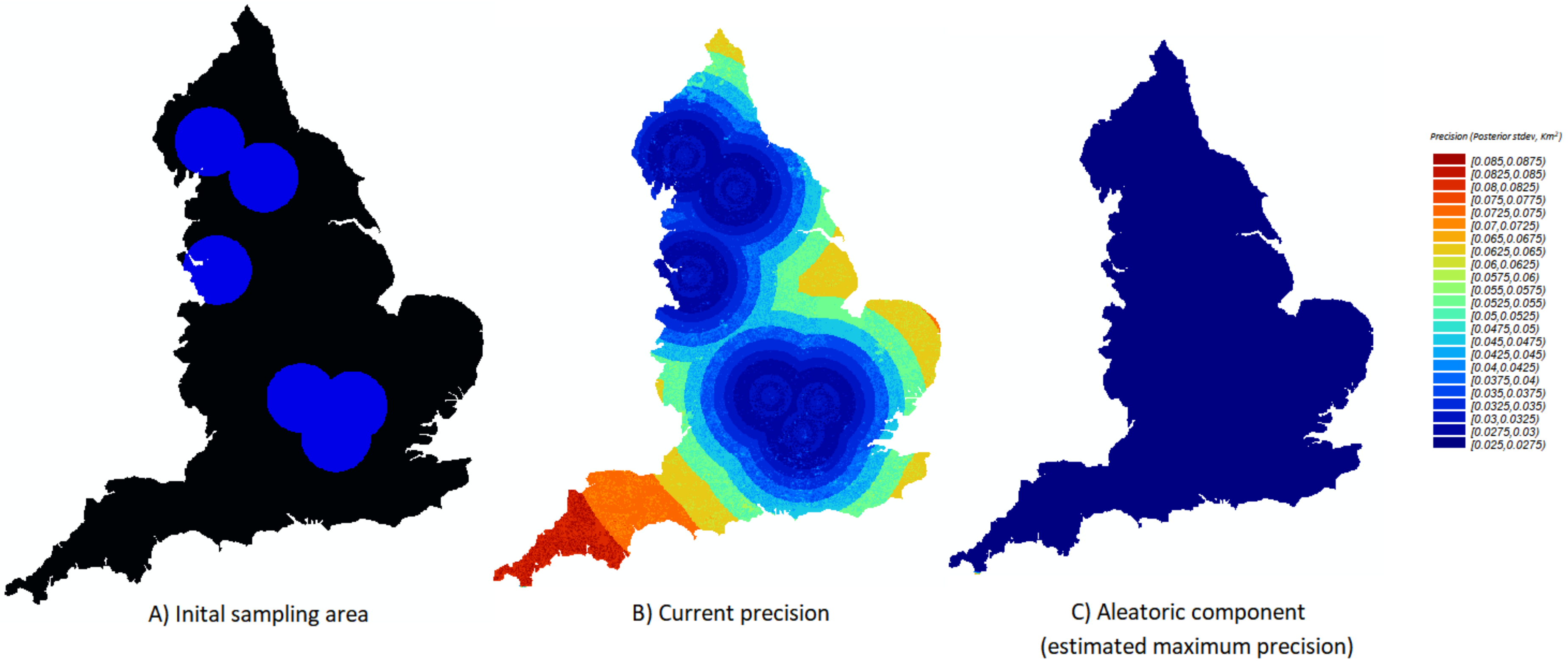


Figure 5.3 *A plot of the kernel regression model described in (5.1) fitted to the initial sample shown in Figure 5.2. By adding the propensity score into the model as a predictor, one can make use of the result in Section 3.2.3 to bypass additional modelling assumptions related to bias in sampling.*

5.3 Using aleatoric and epistemic components to propose targeted sampling designs

With the bias in the initial sample accounted for in the first quantification of uncertainty, the next step is to propose some appropriate sample designs for the next iteration (including the possibility of not sampling further). Much like the second part of the Lagos study, one may view how different components of uncertainty compare to the current uncertainty across the map (Figure 5.4) and across the predictive feature space (Figures 5.5 and 5.6). In this case study though, there are regions where the current uncertainty is much greater than its aleatoric component.

In Figure 5.4, these areas may be seen by comparing 5.4b with 5.4c; where the red areas indicate the largest gap between the current and aleatoric component of uncertainty and the blue areas indicate areas of similarity. When viewing the components across the predictive feature spaces in Figures 5.5 and 5.6, one can see that, as the propensity scores decrease, the current level of precision begins to increase well above the estimated aleatoric component.



A) Inital sampling area

B) Current precision

C) Aleatoric component (estimated maximum precision)

***Figure* 5.4.** *A view of how the current precision and aleatoric components of uncertainty compare spatially. (A) a map of the target area used in the initial sample design. (B) a map of the current level of precision for woodland area predations. (C) a map for the estimated aleatoric component of uncertainty, a measure of the maximum level of precision for predictions under this model.*

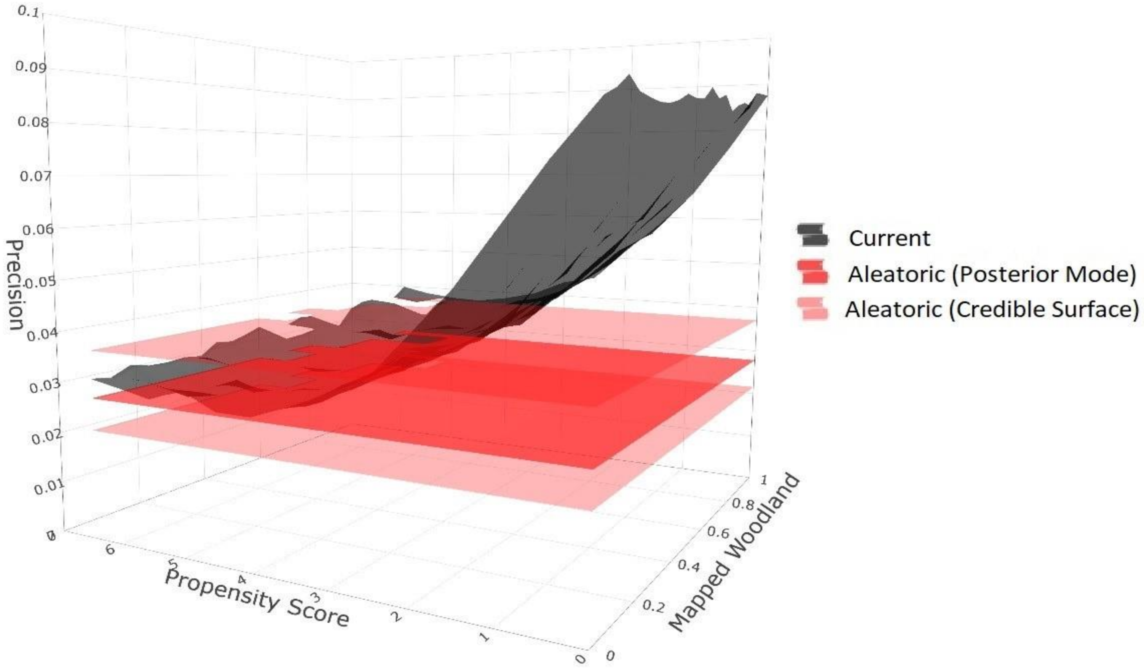


Figure 5.5. *Measures of precision across the predictive features in 3D space (mapped woodland and propensity score). The black surface represents the current level of precision. The red surfaces represent estimates for the aleatoric components (posterior mode and 95% credible surfaces).*

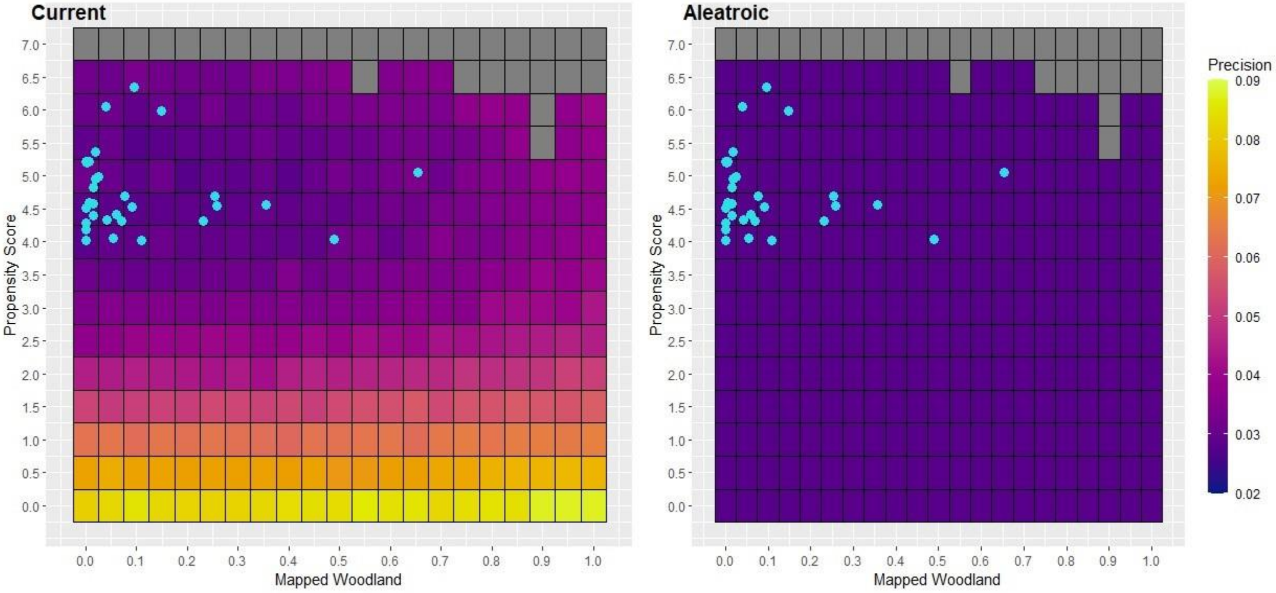


Figure 5.6. *Measures of precision across the predictive features via 2D heat maps. The light-blue points indicate the initial sample.*

Figures 5.4-5.6 all suggest that sampling alone will do little to reduce uncertainty in areas with a high propensity score (i.e. areas close to where the surveyors are based) and that if sampling is to reduce uncertainty in a meaningful way, it will be in areas that are outside of where the initial sample design was conducted. This implies it may be better to go further out when sampling (even if it means reducing the total sample size). At this point though, it is difficult to give explicit formulae for the optimal balance between the relationship between distance from the surveyors' homes, total sample size, and likely reduction of uncertainty.

However, the analysis of these figures is still useful when formulating future sample designs. Figures 5.5 and 5.6 play an important role in generating the sample design for the next iteration, as any probabilistic sample defined in terms of the predicted woodland and propensity score is well suited to make UQ in the next iteration far simpler (see Section 3.2.3 for further details). The main use of Figure 5.4 in this situation is in contextualising any proposed designs set out using Figure 5.5 in terms of spatial clustering and the abundance of different subpopulations. Using Figures 5.4- 5.6 as guides, three designs for further sampling are proposed.

- Design 1 (blue): A larger-sized sample (120) in the same areas as the initial sample (i.e. a propensity score greater than or equal to 4). This design has been selected to examine the hypothesis that there is little to be gained when sampling from this area alone and that venturing out into further areas will be necessary.

- Design 2 (green): A modestly sized sample (20) targeting propensity score greater than 1.8 but less than 2.2. This design has been chosen to consider the possibility of experts visiting further away areas. Because of the COVID restriction on staying overnight, visiting a large number of sites in these areas may not be possible.

- Design 3 (yellow): A modestly sized sample (20) with a mapped woodland area greater than 0.5 and a propensity score greater than 1.8 but less than 2.2. This is similar to design 2, except it also targets areas that have a higher mapped woodland value. This design is chosen to take into account that woodland areas are relatively rare in the mapping and expected to be spatially clustered.

Figure 5.7 shows the targeted areas for each design across the England mapping. Note that, since all three sample designs are defined in terms of propensity scores and the mapped woodland values, one can easily update the posterior distributions using the result from Section 3.2.3.

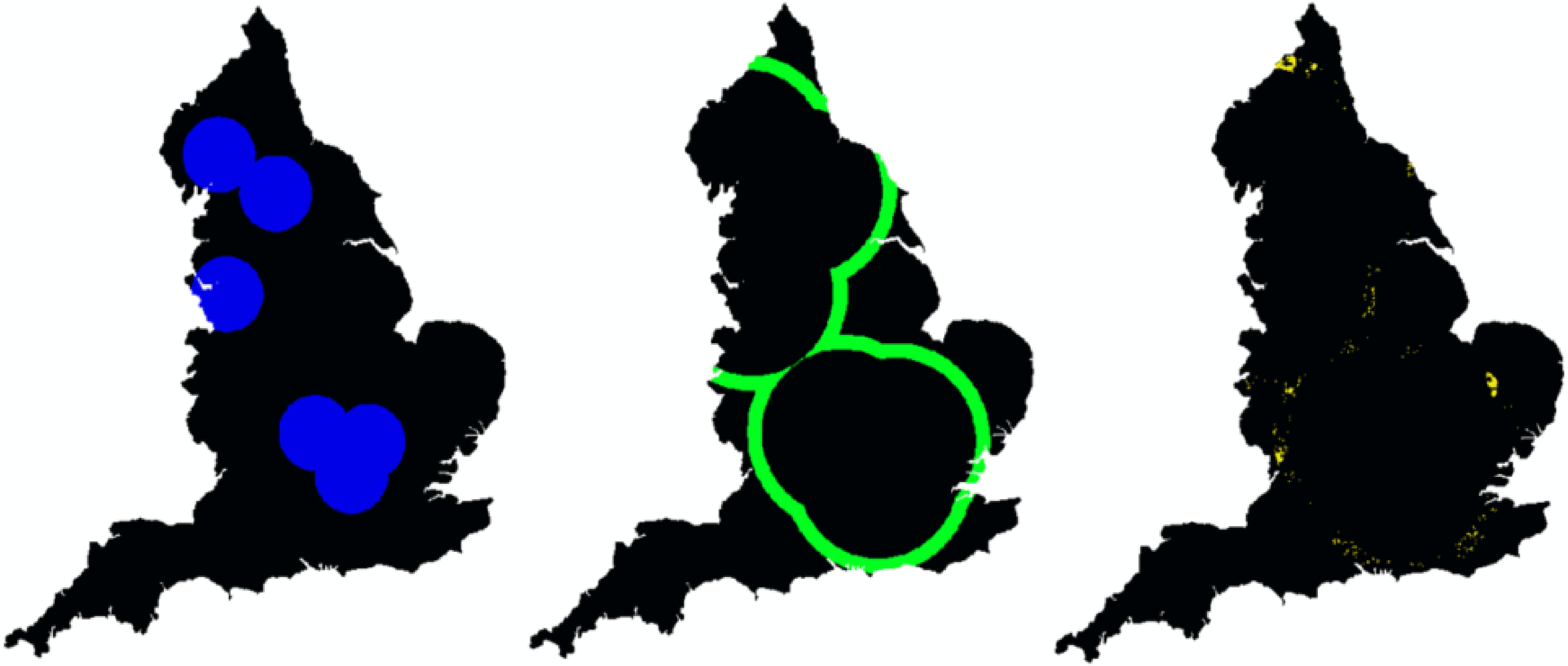


Figure 5.7. *Spatial mappings for the targeted areas under each sample design (design 1: blue, design 2: green, design 3: yellow).*

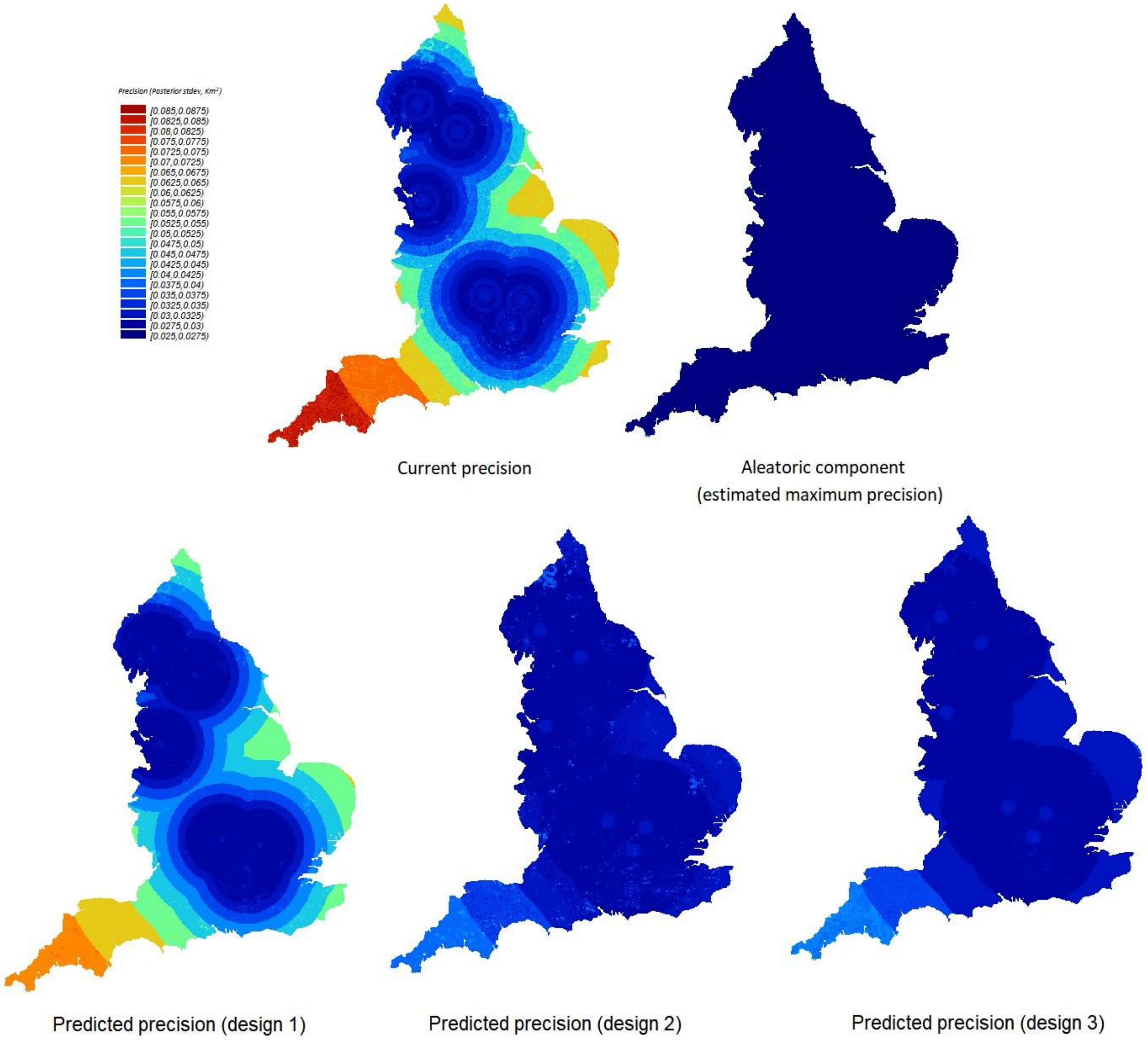
5.4 Predicting the likely effects of the sample designs

With the three sample designs proposed, one reaches the design assessment phase of the adaptive sampling framework. Like the first part of the Lagos case study, one may assess the designs by predicting the likely impacts on uncertainty via a combination of marginalisation and Monte Carlo integration. From Figures 5.7-5.9, the following is observed:

- Sample design 1 is likely to have little impact on the precision of the predictions when compared to the current precision using the initial sample alone. This can be seen throughout figures 5.7-5.9 as the current results are similar to the predicted results for design 1 (i.e. one expects to see little reduction in uncertainty). Sample designs 2 and 3 are likely to be more effective than design 1 for reducing uncertainty in the predictions of the 1km woodland areas.

- The predicted precision under designs 2 and 3 are close to the aleatoric standard deviation for a large area of the map across all the figures. This suggests that for a substantial proportion of the map, there is a good chance that sample designs 2 and 3 will be enough to achieve the maximum possible precision (under this model) for predictions of woodland extent in 1km areas.

- The results for designs 2 and 3 are similar across all three figures. This means it is not clear which will be more effective for increasing the precision of woodland predictions at this stage. In other words, designs 2 and 3 are likely to be as effective as each other for reducing uncertainty based on the initial sample.



Current precision

Aleatoric component (estimated maximum precision)

Predicted precision (design 1)

Predicted precision (design 2)

Predicted precision (design 3)

Figure 5.7. *The predicted precision for woodland area predictions under the three proposed sample designs presented spatially.*

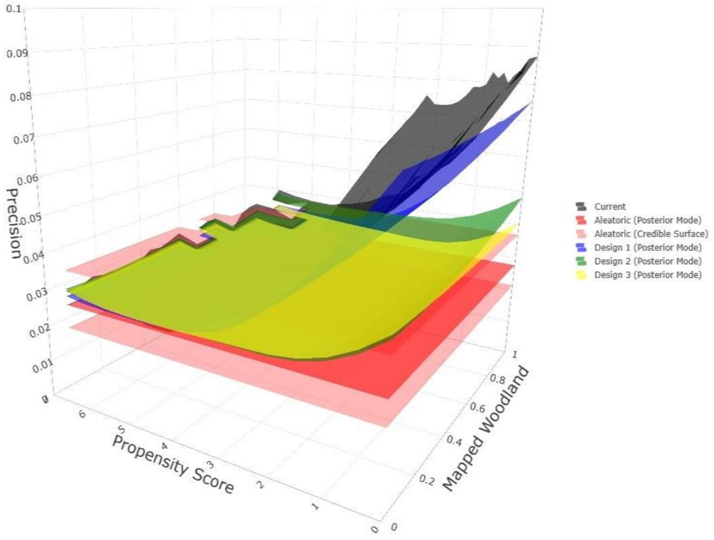


***Figure 5.8**. The predicted precision for woodland area predictions under the three proposed sample designs presented across the predictive features in 3D space (mapped woodland and propensity score).*

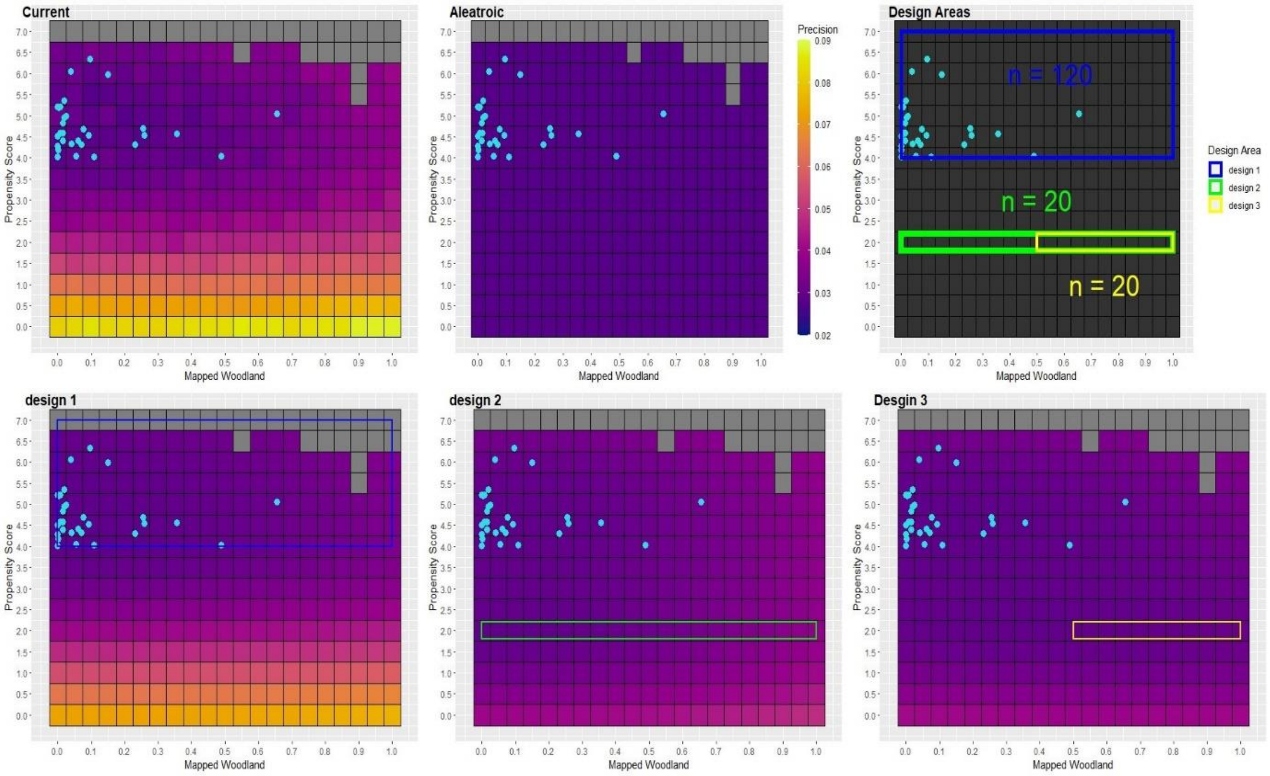


***Figure 5.9.** The predicted precision for woodland area predictions under the three proposed sample designs across the predictive features via heat maps. The light-blue points indicate the initial sample, and the coloured rectangles display the target areas for the proposed sample designs.*

From a decision-making perspective, these observations suggest that firstly, it would be better to venture further away from the experts' homes to apply designs such as 2 or 3, even if it comes at the expense of a smaller sample size. Secondly, they suggest that it may be best to apply sample designs such as 2 or 3 (and then perform a second iteration of adaptive sampling) before committing to designs with larger sizes. This is because there is a strong possibility that the additional reference data from these modestly sized samples will be enough for a substantial proportion of the map. Hence, by applying one of these modestly sized samples first, one can then lessen the risk of wasting resources on unnecessary reference data.

As an aside, it may be difficult to distinguish between designs 2 and 3 based solely on their ability to increase the precision in predictions (i.e. reduce uncertainty) at this stage. However, there may be other factors to consider from a practical perspective. For example, the spatial clustering of design 3 can be convenient when physically visiting areas to obtain ground-truths. On the other hand, the fact that sample design 2 is defined using only the propensity score has the advantage of not requiring the predicted woodland extent when applying the methods in Section 3.2.3. Two examples where this may be desirable are when using the reference data to fit other models (e.g. for other land use classes) or when one wants to allow the predicted values for the woodland extent map to change (e.g. updating the map as new information becomes available or when misclassifications have been recognised).

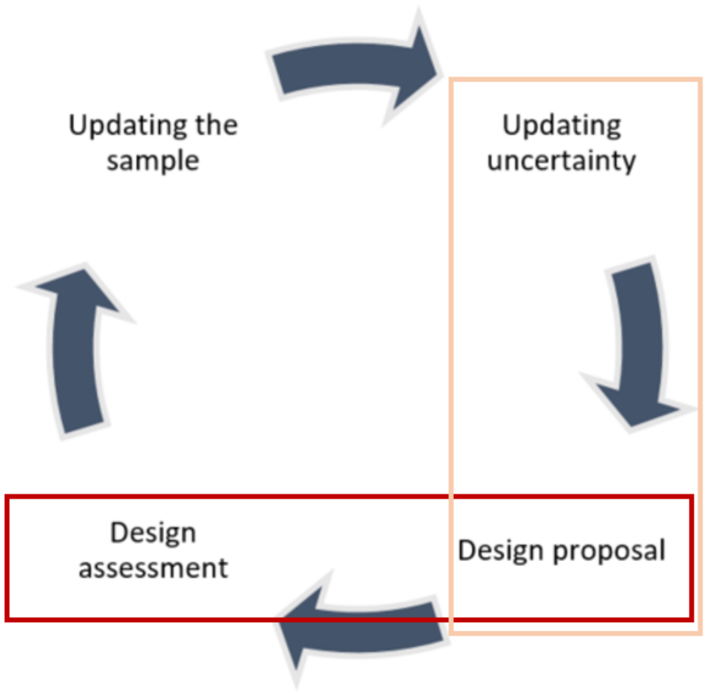
5.5 Reflections (Part III)

Overall, this woodland study has illustrated how the adaptive sampling framework introduced in Chapter 3 can be used to facilitate an adaptive sampling approach when facing design restrictions (see Figure 5.10 for a full summary).

Start: *the initial sample has been collected under targeted sampling defined through a propensity score.*

Objective: *decide upon an appropriate design for the next phase of sampling (total sample size, how to target sampling).*

End: *sample designs 2 or 3 are likley to be good choices in this setting.*



Propensity score determines bias in the initial sample. This may cause issues in UQ if not accounted for.

Updating uncertainty after another sampling iteration is much easier under Bayesian inference.

Including propensity scores as modelling features allows one to account for biased sampling in UQ.

Design restrictions related to propensity scoring makes proposing efficient sample designs more difficult.

Defining targeted sampling through model features makes UQ much simpler in the next iteration.

Quantifying aleatoric components of uncertainty can offer a useful guide for where to focus future sampling.

Predicting the likely effects of different sample designs is vital for assessing the different trade-offs when deciding which design to implement.

Bayesian inference allows one to propagate uncertainty in parameter values easily with Monte Carlo methods when predicting the likely effects of further sampling.

Having the aleatoric component of uncertainty helps contextualise results when predicting the likely effects of further sampling.

Defining targeted sampling through the model features synergises well with Monte Carlo methods.

Uncertainty in parameter values makes proposing and assessing sample designs more difficult.

Methods for suggesting optimal balancing of trade-offs are not readily available.

Aleatoric uncertainty exists and limits the maximum level of precision in modelling.

Figure 5.10. *A summary of how the adaptive sampling framework helped in overcoming the challenges in adaptive sampling for the England woodland mapping case study.*

In particular, the England woodland case study has illustrated that by using the propensity score as a predictor in a model, one was able to include the targeted initial sample without needing to rely on additional assumptions. The idea of using propensity scores in the model then set in motion an iterative process in which one could generate targeted sample designs without needing additional modelling assumptions, so long as the design could be written as a probabilistic sample defined using the propensity score (or some combination of the predictive features and the propensity score).

This advantage to forgo additional modelling assumptions was relied upon many times with other methods of the framework, as it meant that one could make use of third-party software without needing to make any alterations to components such as likelihood functions (which may have required a lot more work).

For example, it is theoretically possible to adjust the model fit from equation (5.1) to account for bias in sample design. However, doing so requires reformulating likelihood functions and developing methods to sample from posterior distributions. This latter task can be time-consuming and require higher expertise (especially if one uses higher dimensional models). By including the propensity score in the model and sample design, one was free to use the third-party software (in this case the bkmr package in R [377] to fit the models and draw from posterior distributions). Consequently, one was able to take advantage of all the quality assurance, optimisation and additional features that have gone into the development of this package. These advantages are heavily compounded in methods such as predicting the likely effects of further sampling, which relies on fitting and applying the same model many times.

The aleatoric components of uncertainty also played a substantial role in this case study. By providing estimates for the maximum level of precision under the current model, one was provided with a useful guide for proposing sample designs in this case study. Furthermore, having estimates for the aleatoric component of uncertainty helped contextualise the results when predicting the likely effects of further sampling under different designs. From here, one was able to establish when some of the proposed sample designs were likely to be *good enough* in the context of the problem. This kind of analysis can become vital in situations when the problem of sample design involves trade-offs between many factors which are not easily expressed in formulaic terms.

Similar to the Lagos case study, the advantages brought by adopting Bayesian inference in UQ were seen throughout the case study. Once again, the concepts of marginalisation and Monte Carlo integration were essential in applying many of the methods.

From a more general perspective, the idea of using propensity scores within models as a means of accounting for purposely biased or targeted sampling can easily be extended to other applications. This is because one can change the factors defining a propensity score without changing the core methodology. For example, one could easily replace the propensity score in the case study with one that uses a more sophisticated model for the accessibility of the areas (e.g. a score that considers the distance from roads, elevation etc.).

Whilst one was able to use the proposed methods to successfully apply adaptive sampling under an example with design restrictions, there were a few limitations in the methods used in this case study:

Firstly, the idea of using propensity scores in models to avoid problems with sample bias requires that one can explicitly state the factors that bias a sample – i.e., the bias is known. This is not an issue if the propensity is pre-defined, as is the case with targeted sampling (e.g. based on known costs or preferences), but this does become an issue when using reference data where the sample design is not strictly controlled or well understood (e.g. relying on volunteered data or found data).

Secondly, there are still major gaps when it comes to proposing efficient sample designs. In this case study, aleatoric components were used to act as a guide for sample designs. Ideally, one would want a procedure that can give explicit recommendations on the sizes of future samples and where designs should be target when facing design restrictions.

Thirdly, Monte Carlo methods can be computationally expensive, and this can become a problem when dealing with higher-resolution imagery or when predicting the effects of many proposed sample designs. In this case study, one was forced to compromise on this by only considering three proposal designs and approximating their effects by considering a grid of discrete points across our feature space. Even with these compromises, predicting the likely effects of further sampling through Monte Carlo simulations took approximately 14 hours for each of the three sample designs on an Intel core i5-8350 CPU.

It should be noted though, that these limitations are not insurmountable when considering the adaptive sampling framework. In fact, there are potential ways around these limitations that fit neatly within the framework. Whilst these potential methods are partially beyond this case study and thesis, a discussion of them is provided in Chapter 7.3.

5.6 Summary

In this chapter, the adaptive sampling framework set out in Chapter 3 was evaluated for a second time, this time on an application involving a woodland mapping of England. This second case study was chosen in response to the first case study as a means of testing the capability of the framework on a more difficult problem. More specifically, this second study differed from the first case study by (i) not having the luxury of a full reference map, (ii) beginning with an initial sample with a known bias, and (iii) having a further restriction to sampling based on propensity scores.

Overall, the England woodland case study illustrated how the framework could successfully incorporate these different challenges to work through an adaptive sampling approach.

Ultimately, the key to the success of the framework in this case study was down to the decision to include the propensity scores in the model that linked the predicted values with their ground-truth counterparts and then to define any targeted sampling through these propensity scores. Once these decisions were set in place, the England woodland study became, functionally, very similar to the Lagos case study. The key difference was that this time, there was enough epistemic uncertainty in the map to make the idea of another round of targeted sampling worth considering. The advantage of using Bayesian inference in UQ was once again highlighted in the England case study, as it meant that simulation-based methods could be exploited to provide estimates for key metrics of the analysis that otherwise may have been unavailable due to a lack of closed-form solutions.

However, despite the success of the framework in this case study, there were still some areas which should be looked at further when considering the framework at a more general level. The first area is that the bias in the initial sample needed to be known and expressed as a probabilistic design based on some known propensity score. In general settings, the factors influencing bias in designs may not be known. The second area is that the sample designs were based on a visual inspection of maps and two- and three-dimensional plots. In a more general classification setting, there may not be a map to project classifications onto and it may be difficult to extend similar plots for higher dimensional settings. Thirdly, whilst the use of simulation-based approaches offers a great advantage in making methods more generalisable, the computational costs were a noticeable burden (e.g. it took 14 hours on an Intel core i5-8350 CPU to predict the effects of further sampling for each design in this case).

With the adaptive sampling framework both applied and evaluated in each of the case studies, the next stage is to reflect on how these methods fit into the wider context of efficiently managing uncertainty in MLTs, and how any methods may be extended or refined to go beyond applications in land cover mappings.

Chapter 6 Evaluation and Discussion

6.1 Introduction

This thesis investigates how one can efficiently manage uncertainty in classifiers in situations where (i) the classifiers may be built using machine learning techniques, and (ii) sampling reference data suitable for uncertainty quantification may be difficult because of practical restrictions such as costs and some members of the population not being easily accessible.

Before starting the investigation, much of the thought was focused on trying to optimise reference sampling in machine learning classifiers to produce an efficient way of reducing uncertainty. After a literature review and some early exploration though, it became clear that an optimisation perspective was too narrow in focus and failed to consider the nuanced (and often subjective) choices that need to be made when balancing uncertainty, sampling restrictions and machine learning.

From these insights, the aim of the thesis remained the same, yet the focus shifted a lot towards a more iterative approach to managing uncertainty that was not overly reliant on specific forms of UQ or MLTs. The motivation for this was that by having an approach to uncertainty management that was agnostic to the choice of MLT and UQ, one would have a more consistent way of managing uncertainty across different applications.

With a focus on a more iterative and generalised approach, Chapter 3 introduced a framework whereby uncertainty could be better managed through adaptive sampling. The construction of this framework began by abstracting the idea of adaptive sampling to a cyclical process and then populating this process with a set of methods and practices.

Chapters 4 and 5 then used two land cover mapping case studies as a way of evaluating this framework by providing challenges that one would expect to see across many adaptive sampling applications (e.g., limited total sample sizes, design restrictions, biased initial samples etc.). However, if this framework is to go beyond these case studies to further applications, it is important to consider how the results from these case studies fit the wider context of both quantifying and managing uncertainty in machine learning.

The purpose of Chapter 6 is to reflect further on the proposed framework with this wider context in mind and will be split into two parts. The first part (Section 6.2) will use the results from Chapters 4 and 5 to evaluate the framework against the criteria set out in Chapter 3. The second part (Section 6.3) discusses a number of important items emanating from this work which include reflecting on the choice of methods within the framework, using the framework to explore trade-offs with design choices, and recognising the interplay between generalisability and efficiency. Section 6.4 then summarises the analysis and thoughts from these two sections.

6.2 Evaluating the framework

6.2.1 Evaluation (I): managing uncertainty efficiently under design restrictions.

An important lesson drawn from this thesis is that there can be restrictions on the data available for UQ (even in situations where data are abundant enough for MLTs to be effective). A key reason for this is that there are often stricter requirements for how the data used in quantifying uncertainty needs to be collected that tend to be punished more by practical limitations.

The design restrictions in Lagos and England case studies are, to an extent, representative of restrictions commonly seen in classification problems. In the Lagos case study, the restrictions focused on the total sample size, which created the problem of deciding how to distribute a limited sample size across the strata efficiently. For the England case study, additional restrictions were placed via propensity scoring, where the initial sample was biased towards where the surveyors were based and sampling far away from these areas would lead to a reduction in the total sample size due to limited travel time.

In the first part of the Lagos study, the framework was used to validate efficient sample designs (which were motivated by non-linear programming) by predicting the likely effect different designs would have on the precision of the area estimate. For the second part of the case study, the framework was used to estimate the limitations of sampling using the aleatoric and epistemic components of uncertainty. This helped manage uncertainty more efficiently by signalling that additional sampling was unlikely to reduce uncertainty further and so it may be best to stop sampling (or possibly consider an alternative model).

The final case study illustrated how the framework can be used to manage uncertainty efficiently. By including the propensity score in the model from the start, one could easily account for the first challenge of bias in the initial sample. Following that, one could use the framework to provide designs that could balance the need for efficient uncertainty reduction whilst also being mindful of the distances needed to travel to locations. This was achieved by using the estimates of aleatoric and epistemic uncertainty to give an idea of which areas were worth sampling and then fine-tuning sample designs by assessing different 'what if?' scenarios.

These case studies have illustrated how the framework can be used to manage uncertainty efficiently under different design restrictions by offering users a way of exploring the potential trade-offs for sample designs between iterations in adaptive sampling. Furthermore, an encouraging feature of this framework is that it can still be applied when facing the additional challenges that come from quantifying uncertainty in machine-learning classifiers.

One limitation of the framework though is that it does not offer explicit ways of recommending efficient sample designs in its current form. That is, whilst the framework does allow one to easily explore different sample designs to eventually make a decision based on efficiency, the user is still left needing to specify these efficient designs in the first place. However, it is important to note that the framework not offering explicit ways of recommending designs may not be the major limitation it first seems. Firstly, the ability to recommend sample designs (with optimisation methods or otherwise) is not a requirement for the adaptive sampling framework to be effective. Secondly, for the sake of generalisability, it may be better to keep methods of recommending sample designs and the framework as separate entities, with the former being used to enhance the latter in specific cases (this is discussed further in Section 6.3.3).

6.2.2 Evaluation (II): generalisability

Another important lesson from the literature review is that managing uncertainty efficiently in machine learning classifiers often comes down to making suitable choices at the intersection of machine learning, uncertainty quantification and methods in reference sampling. Given the amount of work already on these topics, the subjective elements in UQ, and the additional challenges brought about by using MLTs, the idea of trying to create a system that provides the best combination of sample design, method of UQ, and MLTs at once does not seem feasible.

Instead, this thesis sought to propose an adaptive sampling framework that could remove the choices of UQ and MLTs from this equation by having a framework that is agnostic to these two choices. This desire to be agnostic to the choice of UQ and MLT was captured via the criterion of generalisability.

For all three parts of the case studies, the stages of creating the maps (which were all done using machine learning classifiers) and quantifying uncertainty were kept separate. Hence, the MLTs in each case would have been free to vary. For the method of UQ, the case studies illustrate how the framework could be used on (i) population-level estimates from discrete classifiers and stratified random sampling (Lagos study, part I) (ii) individual-level estimates based on fuzzy classifiers and modelling (Lagos study, part II) (iii) individual-level estimates based on fuzzy classifiers through modelling with biased sample designs influenced by a known propensity score (England woodland study). In these cases, the structures of the models that linked the predicted values to the ground-truths were also free to vary, along with how the strata were defined in the area estimation part of the Lagos case study and how the propensity scoring was defined in the woodland case study. Furthermore, replacing any of these components would not have changed the fundamental workflow thanks to the combination of Bayesian inference and Monte Carlo methods.

Consequently, the framework offers a high degree of generalisability. In its current form, there may be limitations when using Monte Carlo methods under computationally intensive models, insisting on UQ under a frequentist perspective, or when experimenting with sample designs in high-dimensional settings. However, these are not major limitations and may better be described as challenges within general modelling rather than a weakness of the framework itself (e.g. high-dimensional modelling comes with challenges with or without adaptive sampling).

6.2.3 Evaluation: overview

Overall, the case studies in this thesis have illustrated how the proposed adaptive sampling framework can offer a generalisable approach for managing uncertainty efficiently under design restrictions. Table 6.1 summarises how the framework has benefited each of the case studies individually, along with any limitations and areas for improvement.

From a more general perspective, the results from these case studies are encouraging. A common thread throughout the studies was that the methods in the framework were agnostic to the choice of classification method, the models used to quantify uncertainty, and propensity scoring used to define which areas were easier to sample from. Hence, it is likely that the framework can be applied to a wide range of applications.

Whilst there are some minor limitations when dealing with more advanced modelling and computational costs, the framework offers a substantial contribution in the aim toward a generalisable approach to managing uncertainty in machine learning classifiers.

***Table 6.1** A summary of how the adaptive sampling framework performed for the Lagos and England case studies against the criteria set out in Chapter 3. Entries with a (+) indicate overall positive or successful features and entries with a (-) indicate limitations or areas for improvement.*

Case Studies	The ability to manage uncertainty under design restrictions.	Generalisability
Lagos urban area estimation (part I) Properties Area estimation problem. A discrete classifier was used. UQ is based on proportion estimates under stratified random sampling. Design restrictions were based on the total sample size.	(+) With the framework, it was possible to conduct a cost-benefit analysis for different design proposals based on the initial sample (with appropriate uncertainty quantification). (+) By experimenting with different design proposals, it was possible to find efficient sample designs.	(+) The framework can be easily applied to other population-level estimates (e.g. overall accuracy, false positive rates, false negative rates etc.). (+) No assumptions were made about the classifier used to produce the original map. (-) Requires Bayesian inference for some important steps, which excludes model-assisted estimators (a popular method based on frequentist inference).
Lagos urban mapping (part II) Properties Mapping problem (individual cases). A fuzzy classifier was used. Model-based UQ under stratified random sampling. Design restrictions were based on the total sample size.	(+) Estimates for the aleatoric component of uncertainty across the map indicated that enough sampling had already been done.	(+) No assumptions were made about the classifier used to produce the original map.
England woodland mapping Properties Mapping problem (individual cases). A fuzzy classifier was used. Model-based UQ under probabilistic sampling. Design restrictions were based on the total sample size and a preference to stay close to where the surveyors were based.	(+) Bias in the initial sample could be easily accounted for by including the propensity score in the model used for UQ. (+) With the framework, it was possible to conduct a cost-benefit analysis for different design proposals based on the initial sample (with appropriate uncertainty quantification). (+) Defining targeted sampling through the predictive features offered a simple way of quantifying uncertainty under targeted sampling. (+) By experimenting with different design proposals and contextualising them with the aleatoric component across the map, it was possible to find efficient sample designs.	(+) No assumptions were made about the classifier used to produce the original map. (+) No assumptions were made about how the propensity score was defined. (-) Accounting for bias in the initial design meant knowing what influenced the bias. This may not be the case with uncontrolled sampling or found data. (-) Some Monte Carlo methods were computationally intensive. This may limit the types of models that can be used when applied to a larger scale. (-) Designs were generated using experimentation and visualising results across maps and the model's feature space. In general, one may not have access to such visualisations.

6.3 Discussion

This section discusses three important items emanating from the work in this thesis, specifically: understanding the role of the methods in the framework (6.3.1); using the framework to experiment with designs (6.3.2); and the interplay between generalisability and the ability to manage uncertainty efficiently (6.3.3).

6.3.1 Reflecting on the choice of methods within the framework.

The adaptive sampling framework in this thesis was presented alongside a set of methods designed to help users navigate various stages of the adaptive sampling cycle.

To better understand why the framework is likely to be successful in other applications (and address any potential limitations), it is important to understand how these methods contribute to an adaptive sampling framework. Table 6.2 provides an overview of how each method contributes individually to making adaptive sampling either more efficient or generalisable.

Table 6.2. *A summary of how the methods introduced alongside the adaptive sampling framework contribute to meeting the criteria set out in Chapter 3. Entries with a (+) indicate overall positive or successful features and end entries with a (-) indicate limitations or areas for improvement.*

	The ability to manage uncertainty efficiently under design restrictions	**Generalisability**
Bayesian inference in uncertainty quantification	(+) Allows prior knowledge to be formally incorporated into UQ. (-) Does little on its own to help users generate efficient sample designs.	(+) Greatly improves generalisability by making UQ under sequential sampling easier and enabling Monte Carlo methods. (+) Vital for improving the generalisability of other methods with Monte Carlo methods. (-) Computational costs can be substantial in high-dimensional models which potentially limits what models could be used in practice. (-) Excludes the use of UQ based on frequentist methods.
Using the predictive features to define targeted sampling.	(+) Offers a simple and scalable approach to combining multiple iterations of targeted sampling. (+) Offers a simple method of accounting for bias in designs through propensity scoring. (-) Does little on its own to help manage uncertainty efficiently under design restrictions.	(+) No major restrictions are placed on the structure of the model. (+) No restrictions are placed on how the propensity scores are generated. (+) Allows one to include biased samples without needing to explicitly model the sample design.
Quantifying epistemic and aleatoric components of uncertainty.	(+) Can improve sampling efficiency by helping users identify when it is time to stop sampling. (+) Can act as a guide to help inform users where further designs should target. (-) Does little on its own to generate explicit rules for optimising sample designs.	(+) Highly generalisable when combined with Bayesian inference and Monte Carlo methods. (+) Easy to apply to other measures of precision.
Predicting the likely effects of further sampling	(+) A vital method in adaptive sampling as it allows one to assess proposed sample designs before they are implemented. (-) Does little on its own to generate explicit rules for optimising sample designs.	(+) Highly generalisable when combined with Bayesian inference and Monte Carlo methods.

From Table 6.2, four important insights arise:

Bayesian inference greatly improves the generalisability of the framework by making other methods easier to apply.

The key advantage of Bayesian inference in UQ is that it indirectly assists in managing uncertainty efficiently and in generalisability by making other methods in the framework a lot easier to apply.

This advantage stems from two key properties. The first property is the ability to exploit marginalisation and Monte Carlo methods, which were used across all three parts of the case studies. This two-part combination was relied upon when predicting the likely effects of proposed sample designs when analytical solutions were not available. Similarly, the ability to exploit marginalisation and MC methods were used to estimate the aleatoric components of uncertainty in the second part of the Lagos study and the England woodland study.

The second key property is that Bayesian inference is naturally suited to sequential sampling in UQ. This is what allows one to easily merge multiple iterations of sampling in UQ. Concerning specific methods in the framework, the ability to easily handle sequential sampling makes predicting the likely effects of sampling far easier when using MC methods. This is because predicting the likely effects of sample designs with MC methods effectively involves updating models with hypothetical sets of data drawn from a simulated design. Here, it does not matter that the data happens to be hypothetical, the same principles that make UQ easier under real sequential sampling still apply.

Using the predictors to define targeted sampling offers a simple way to include biased and targeted sampling.

Using the predictive features for targeted sampling is a helpful tool for simplifying many of the key steps in the adaptive sampling framework.

To understand why this is the case, one can consider the woodland mapping example. Here, the bias in the initial sample could be easily dealt with by including the propensity score as a feature in the model used to quantify uncertainty. Following that, one could propose sample designs that were targeted to improve efficiency yet did not require further modelling assumptions or alterations. This advantage was especially relevant in the England case study, as one relied on third-party software to generate posterior distributions for the aleatoric components of uncertainty and likely effects of further sampling, which is hard to alter in practice without substantial investment.

At a more general level, using predictors for targeted sampling is not a requirement for adaptive sampling, nor will it guarantee optimal uncertainty reduction. However, it does offer the assurance that such designs will seamlessly fit into any subsequent analysis and UQ steps. Given the difficulties of including targeted or biased sampling in UQ in the general setting (which are only further enhanced when using MLTs, see Chapter 2 for further details), this assurance makes the idea of searching for efficient sample designs in the predictive feature space appealing from an adaptive sampling perspective.

Components of uncertainty are useful heuristics when managing uncertainty.

The woodland case study and urban mapping part of the Lagos case study both relied upon estimates for aleatoric (and by extension epistemic) components of uncertainty. In the Lagos study, the aleatoric component of uncertainty indicated that further sampling was not necessary across the entirety of the map. In the woodland case study, the aleatoric component of uncertainty indicated that there was little to be gained in continuing to sample close to where the surveyors were based and offered a way to contextualise design proposals when predicting the likely effects of further sampling. Hence, these case studies illustrate how estimating the aleatoric components of uncertainty is a useful tool when trying to manage uncertainty under design restrictions.

However, there are some limitations to using the components of uncertainty to manage uncertainty. Firstly, the aleatoric and epistemic components of uncertainty at this stage can only offer a guide for sample design. Ideally, one would want a way of automatically recommending sample designs based on these components as experimenting with different design types is not always easy in high-dimensional settings. Secondly, there are situations where differentiating between aleatoric and epistemic uncertainty offers little practical use in managing uncertainty efficiently. The first part of the Lagos case study is an example of such a case, as there was technically no aleatoric component of uncertainty. Thirdly, the issue of quantifying the ontological component of uncertainty is left unresolved, meaning that one must implicitly assume that any assumptions in UQ are valid beforehand when quantifying components of uncertainty. Section 7.3.1 discusses how future work could incorporate ontological uncertainty through adaptive modelling.

Predicting the likely effects of sample designs offers a universal and evidence-based approach to assessment between iterations in adaptive sampling.

For both the estimation problem in the Lagos study and the woodland mapping problem, a large part of deciding which design would be best to adopt in the next iteration was based on how different designs were likely to impact uncertainty based on the current information. The advantages of this approach to design assessment lie in how universal it is and how it can incorporate sources of uncertainty when combined with Bayesian inference and iterative sampling.

The choice between sample designs in the England case study is a good illustration of these advantages. Here, all the designs were generated by experimenting with the balance between the distance from where the surveyors are based with the likely uncertainty reduction. None of the designs were generated by trying to optimise some objective function and to do so would have been difficult to justify, as factors such as spatial clustering and the true cost of travel may not align neatly with an objective function that could be easily optimised.

Nevertheless, one could predict the impact of each design and make the case that designs 2 and 3 (which opted to go for fewer surveys further away) were likely to be better for uncertainty reduction than design 1 (which focused surveying on a larger number of sites that were close to where the surveyors were based).

As a side note, the ability to assess the likely impacts of *any* design is what allows one to start employing an evidence-based approach to exploring different trade-offs in adaptive sampling, an important topic discussed in Section 6.2.3.

6.3.2 Using the framework to explore different design choices in adaptive sampling

A recurring theme throughout this thesis is that creating sample designs to optimise uncertainty reduction in machine learning classifiers can be difficult. Two major causes of these difficulties are (i) needing sufficient knowledge of a system beforehand (e.g. having a good idea of suitable model structures and parameter values) and (ii) cost-benefit trade-offs of sampling decisions not falling into well-defined objective functions. Because of these difficulties, this thesis decided to take an alternative approach to balancing uncertainty reduction and sampling restrictions in machine learning techniques, proposing that uncertainty could be better managed via adaptive sampling. With this change in perspective, Chapter 3 proposed a framework with a series of methods that could enable such adaptive sampling.

After reflecting on the case studies, one of the key advantages this framework offers is its ability to provide an easy way of exploring different sample designs based on the data from previous iterations. With this ability, the two causes of optimisation difficulties that were previously mentioned do not have the same degree of impact. In the case of needing sufficient knowledge beforehand, the framework mitigates this issue by representing the impact of different design choices as probabilistic statements based on the current data. This means that there is less concern over making the correct assumptions at the beginning of the sampling process, as uncertainty within a system can be accounted for and updated between iterations of sampling. As for situations where the cost-benefit trade-offs cannot be easily represented through objective functions, the ability to propose and implement different designs lets one consider trade-offs at higher levels.

In short, with the proposed framework, one can easily explore targeted designs until one finds a suitable design. As situations involving uncertainty management become more complex, this approach based on experimentation becomes far easier to apply consistently when compared to an optimisation perspective.

This is not to say that optimisation methods cannot play a useful role in adaptive sampling. Instead, it may be better to view optimisation procedures as methods that fit into the design proposal stage of the framework. From this perspective, one may still use optimisation methods to suggest sample designs, but they are treated as any other design created through experimentation.

An example of using optimisation methods to complement the framework can be seen in the first part of the Lagos study. Here, design (ii) was motivated by a non-linear programming problem that was optimised on the assumption that true values of the relevant parameters were equal to the modes of their respective posterior distributions. However, because one was able to use the framework to predict the likely effects of further sampling, the validity of design (ii) was no longer beholden to these assumptions being true. Instead, the framework could compare designs without requiring a formal justification based on assumed values.

6.3.3 Recognising the interplay between generalisability and efficiency.

The final discussion topic is that there is an interplay between managing uncertainty under design restrictions and generalisability. Figure 6.3 illustrates this idea but, briefly, as a framework becomes more generalisable, one has more options available when managing uncertainty, which increases the chances of finding efficient ways of reducing uncertainty. Conversely, being able to manage uncertainty efficiently under different design restrictions will inevitably make some previously unviable methods practicable again, hence increasing generalisability.

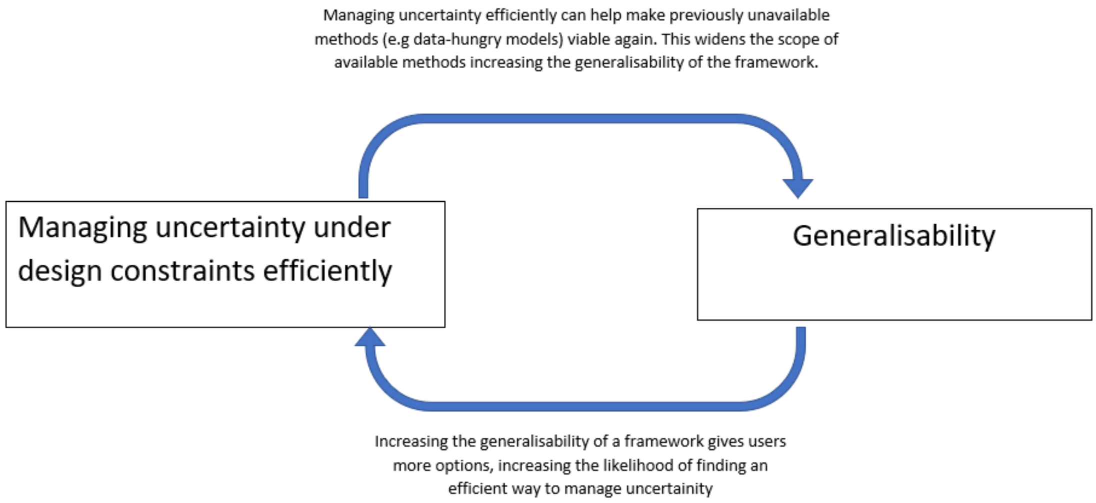

Managing uncertainty efficiently can help make previously unavailable methods (e.g data-hungry models) viable again. This widens the scope of available methods increasing the generalisability of the framework.
Managing uncertainty under design constraints efficiently
Generalisability
Increasing the generalisability of a framework gives users more options, increasing the likelihood of finding an efficient way to manage uncertainity

Figure 6.3 *A summary of how generalisability and efficiency can feed into each other to illustrate that these topics should not be viewed in isolation.*

Recognising this interplay is an important part of understanding the strengths of the framework and shaping further development as the underdevelopment of one area will inevitably lead to limitations in the other.

The woodland mapping case study illustrates this interplay. With a generalisable way of experimenting with different targeted sampling designs, it was possible to create at least two efficient sample designs. This is an example of generalisability feeding into managing uncertainty efficiently. An example of managing uncertainty efficiently feeding into generalisability comes at beginning of the all the case studies. To recall, the core problem of all the case studies was that the maps were generated with machine learning classifiers which were assumed to be too black-box to quantify uncertainty directly. Instead, the case studies chose to use UQ based on a separate reference sample. Because of practical restrictions though, the designs of these reference samples need to be carefully considered. Thankfully, the adaptive sampling framework could help in managing uncertainty efficiently under the design restrictions. These results ultimately feed back into generalisability, as it clears a major bottleneck in UQ under maps made with MLTs. In other words, by having a way of managing uncertainty efficiently for UQ that does not rely on the choice of MLT, one inevitably has more classification methods available (hence increasing generalisability).

6.4 Summary

This chapter has provided an evaluation of the proposed adaptive sampling framework and raised important issues for discussion emanating from the work.

For the evaluation, the adaptive sampling framework largely met the two criteria by offering a generalisable way of managing uncertainty under design restrictions. The case studies illustrate how the framework can be used to manage uncertainty under design restrictions such as limits on the total sample sizes when the cost of sampling varies across a population. In addition, the case studies illustrated a high degree of generalisability for the framework as many of the methods were agnostic to the choice of classification method, the model used to quantify uncertainty, and propensity scoring. Whilst there are some potential limitations when dealing with computationally intensive models and in high-dimensional modelling, the framework as a whole offers a substantial contribution towards a generalisable approach to managing uncertainty in machine learning classifiers.

The discussion section focused on three main topics: (i) reflecting on the choice of methods within the framework. (ii) using the framework to explore different design choices in adaptive sampling (iii) recognising the interplay between generalisability and efficiency.
From the first part of this discussion, the following points emerge:

- Bayesian inference greatly improves the generalisability of the framework by making other methods easier to apply.
- Using the predictive features to define targeted sampling offers a simple way to include biased and targeted sampling.
- Components of uncertainty are useful heuristics when managing uncertainty.
- Predicting the likely effects of sample designs offers a universal and evidence-based approach to assessment between iterations in adaptive sampling.

The second discussion topic was that the adaptive sampling framework is at its best when it is used as a tool for exploring trade-offs between uncertainty reduction and design restrictions rather than trying to use it as a means of optimising sample designs. The argument here begins with the realisation that optimising sample designs is an unrealistic prospect once one considers that sample design itself is often subject to uncertainty and those design restrictions do not always follow well-defined objective functions. Instead, it can often be better (perhaps with the aid of optimisation methods) to use the framework to experiment with different designs and explore different trade-offs. In short, where the first topic in this discussion focused on how the methods allow one to easily experiment with different designs, this topic discusses that such an explorative-based approach is a better way of viewing the problem of uncertainty management under design restrictions in the first place.

The final topic focused on the observation that the original criteria (the ability to manage uncertainty efficiently under design restrictions and generalisability) are not separate but interlinked. This is an important lesson for future applications as it offers new perspectives when facing the problem of managing uncertainty under design restrictions and highlights that managing uncertainty effectively requires a sufficient degree of generalisability and vice-versa.

Chapter 7: Conclusion

7.1 Summary of the thesis

Classification via machine learning has become increasingly popular across a myriad of domains, especially in the last few decades. This popularity is largely due to their ability to automatically (or at least with a high degree of automatability) produce high-quality models from large sets of data. As these techniques have gained in popularity and spread into more applications, there has been increasing demand to quantify uncertainty to the same standard one had typically reserved for more traditional forms of modelling such as process-driven modelling or data-driven modelling in lower-dimensional spaces.

However, because many methods of machine learning are black-box in nature, rely on large quantities of data, and have not always been designed with uncertainty quantification in mind from the outset, quantifying uncertainty often relies on having the ‘right’ kind of reference data. These ‘right’ kind of reference data typically involves higher quality data (e.g., ground-truth observations) collected under strictly defined sample designs. Ultimately, this tends to create situations where machine learning techniques may be held back in applications that demand uncertainty quantification due to a lack of this higher-quality data collected under suitable sampling conditions.

This thesis has investigated an adaptive approach to sampling as a means of efficiently managing uncertainty in classification where (i) elements of the classifier rely on modern machine learning techniques and, (ii) the amount and types of suitable reference data are limited in some way by design constraints. This was achieved through the following objectives:

A. Developing a framework for adaptive sampling that allows users to efficiently manage uncertainty in classifiers built with machine learning techniques. This framework should allow users to leverage the information contained in an initial reference sample to make informed and more cost-effective choices related to further sample designs when quantifying uncertainty under classifiers built using machine learning techniques.

B. Evaluating the proposed framework using a series of land cover mapping applications.

C. Providing recommendations on how the proposed framework could be further developed to address any unresolved weaknesses found in the evaluation stages.

To meet objective A, this thesis began with an extensive literature review aimed at the topics of uncertainty quantification, machine learning, and methods for generating efficient sample designs for reference data. From the review, it became clear that there was a substantial amount of work that covered all these topics individually, as well as their pair-wise intersections. Despite this though, there were still several challenges (and a noticeable gap in the literature on how to solve them) when considering all three topics simultaneously.

The major causes of these challenges were typically a result of the following five factors conflicting with each other:

(i) Many MLTs are not designed with UQ in mind and are commonly black-box in nature.

(ii) MLTs often use large data sets collected under unstructured sample designs.

(iii) Efficient uncertainty reduction is heavily influenced by the performance of the classifier (whether built with MLTs or otherwise).

(iv) There are many subjective components in UQ that are typically related to what kinds of modelling and design assumptions one is willing to accept.

(v) The effects of different sample designs are themselves subject to uncertainty as they often rely on unknown values.

It was from this point that the author proposed that any framework of adaptive sampling focused on uncertainty reduction and machine learning classifiers should be as agnostic to specific methods of UQ and MLTs as possible. The motivation for this decision begins by noting that in many ways, the circumstances of the task and the subjective beliefs of stakeholders will tend to dictate what MLTs and methods of UQ will be used in the end. Furthermore, the choices for these methods are likely to change (possibly quite a lot) over time. Hence, it seemed more worthwhile to develop a framework of adaptive sampling where one had a high degree of flexibility over these choices rather than trying to optimise sample design for specific sets of circumstances.

Under this change of mindset, this thesis then proposed a generalisable framework for adaptive sampling that focused on being agnostic to choices within machine learning and uncertainty quantification. This began by breaking the adaptive sampling into a cyclical process consisting of four key stages and then populating this framework with a series of methods and analytical tools that were designed to be as generalisable as possible. Under this framework, the four key stages were *updating the sample, updating uncertainty*, *design assessment*, and *design proposal* and the methods/ analytical tools considered were: the use of Bayesian inference in UQ, using the predictive features to define targeted sampling, quantifying epistemic and aleatoric components of uncertainty, and predicting the likely effects of further sampling.

The thesis then moved on to meeting objective B by evaluating the framework and proposed methods in two case studies spread over three parts. The first case study used an urban mapping problem which was split into two parts, an area estimation problem involving discrete classifications (part I) and a mapping problem using fuzzy classification (part II). The second case study involved a woodland mapping problem with additional design restrictions that favoured sampling areas close to the surveyors' homes as a way of overcoming COVID-19 travel restrictions.

In the first part of the Lagos case study, the framework was successfully used to compare the likely impacts different forms of stratified sampling would have on the precision of area estimates. This offered a great advantage when it came to deciding upon a single design to use in the next iteration of sampling, as it gave a way for selecting between sample designs without needing to implement them.

For the second part of the Lagos study, the framework was used to manage uncertainty efficiently by providing a way of knowing when it is best to stop sampling. This was done by quantifying epistemic and aleatoric components of uncertainty and then concluding that the current uncertainty was so close to the aleatoric component that it was highly likely that further sampling alone would do little to increase the precision of estimates.

Together, both parts of the Lagos case study highlighted how being able to predict the likely effects of different sample designs and the associated maximum level of precision are vital tools when trying to manage uncertainty efficiently through adaptive sampling.

The England woodland study built upon the successes of the framework in the Lagos study and sought to test the framework further with the addition of a biased initial sample and propensity scoring. Under these conditions, using the predicted features to define targeted sampling provided a way of accounting for initial sample bias and a basis for targeted sampling that did not rely on any additional modelling assumptions. The lack of added modelling assumptions had the additional benefit that it made refitting models using third-party software much easier, as it meant one could use many of their inbuilt features (e.g., methods for generating posterior distributions) without needing to make adjustments.

This second case study illustrated how the framework and proposed methods could come together to provide an evidence-based means of managing uncertainty efficiently. In particular, one was able to use the framework to show that it would probably be better to conduct fewer surveys in less favourable areas than it was to continue surveying more sites in highly favourable areas.

Finally, the thesis moved on to objective C by reflecting on how the framework and proposed methods therein could generalise to different kinds of classification problems and areas for further development. Focusing on the proposed methods specifically first, the stand-out results from this thesis are:

- Although not strictly required in the framework, choosing to adopt Bayesian inference in uncertainty quantification is highly recommended as it is naturally suited for sequential sampling and facilitates a generalisable way of applying other elements of the framework thesis via Monte Carlo methods.

- Using the predicted features of a model as a basis for targeted sampling offers users a way of targeted sampling that does not require any additional modelling assumptions. This property is especially useful when there are multiple iterations of targeted sampling because if each design sample in a chain has this property, then so does its composition. Furthermore, this method can be used as an easier way to account for sample bias by forcing the factors influencing bias into the model to begin with.

- Quantifying aleatoric and epistemic components of uncertainty is a useful tool for knowing when one has reached the limits of further sampling and can offer a useful guide for targeted sampling. However, more work is needed to investigate how to best leverage these components of uncertainty when simple visual inspections are not available (e.g., high dimensional models).

- Predicting the likely effects different sample designs will have on uncertainty is an almost essential tool in adaptive sampling, as it serves as an evidence-based method for deciding which designs to implement in the next iteration. In general, this type of approach should be used over benchmarking methods that test the performance of a design after it has been implemented. It should be noted though, that predicting the likely effects of different sample designs can become computationally expensive when the initial model is complicated and one is relying on Monte Carlo methods.

From a more holistic perspective, the case studies illustrate how the methods used to populate the proposed framework can help at different stages of the adaptive sampling cycle and when combined can offer a generalisable approach to adaptive sampling. A second key lesson is that a large component of adaptive sampling lies in experimenting with sample designs under uncertainty rather than trying to optimise sample designs based on limited information. A final lesson is that there is an interplay between managing uncertainty efficiently and generalisability which means that a lack of one places limits on the other. In other words, it can become difficult to manage uncertainty efficiently without a sufficient degree of generalisability. Likewise, there will naturally be a limitation in the kinds of MLTs, and methods of UQ one can employ (i.e., a lack of generalisability) if one cannot manage uncertainty efficiently for these methods.

7.2 Research contributions

The work in this thesis has made the following contributions towards a more cost-effective way of managing uncertainty in machine learning techniques under design restrictions:

The introduction of a strategic level framework for adaptive sampling that is agnostic to the choice of machine learning techniques and type of uncertainty quantification.

One of the major insights from this thesis is that there are many viable approaches to machine learning and uncertainty quantification. The right combination of these methods will often be context-specific and contain many subjective choices (e.g. deciding which modelling assumptions are appropriate).

Considering this, this thesis produced a strategic-level framework for adaptive sampling that provides a way to manage trade-offs between sampling restrictions and uncertainty reduction without relying on specific forms of machine learning techniques or uncertainty quantification. The fact that the framework is agnostic to the choice of machine learning and uncertainty quantification approaches means that the framework can be applied without getting into the previous discussion over the most suitable methods. This property is especially useful for machine learning applications, where it can be hard to control or understand every process involved when compared to more traditional modelling.

The development of tactical-level methods for the adaptive sampling framework

With a framework adaptive sampling set at the strategic level, the thesis then proposed a series of methods that were designed to help users navigate the framework's key stages at the tactical level. These methods were designed to help users answer key questions such as "*How should uncertainty be quantified when new data becomes available?*", "*Where are the best places to sample from under design restrictions?*", "*How do I know when to stop sampling?*", and "*How do I decide on a sample design given several viable candidates?*" Individually, these methods help overcome many of the key obstacles one is likely to face in adaptive sampling. Combined though, these methods offer a generalisable way of completing full iterations of the adaptive sampling cycle and form a strong foundation for future work.

An illustration of how iterative and explorative approaches to sampling design can offer a consistent and flexible way of managing uncertainty efficiently.

When trying to manage uncertainty efficiently, there is always a balancing act when deciding how reference data should be collected between the potential uncertainty reduction and staying within context-specific design restrictions. From the work in this thesis, it becomes clear that managing this balancing act can be challenging in practice because of uncertainty within the system itself (e.g. uncertainty in model choice, parameter values etc) and the fact that design restrictions do not always follow neatly defined objective functions.

With the use of the case studies, this thesis has been able to show how the adaptive sampling framework can be used to provide an iterative and explorative approach to sample design and has illustrated how this can be used to address these challenges. From this, one can see how the iterative component here helps the framework be consistent by building up to efficient sample designs with evidence from the previous iterations. With the explorative component, the framework can remain flexible to different forms of design restrictions.

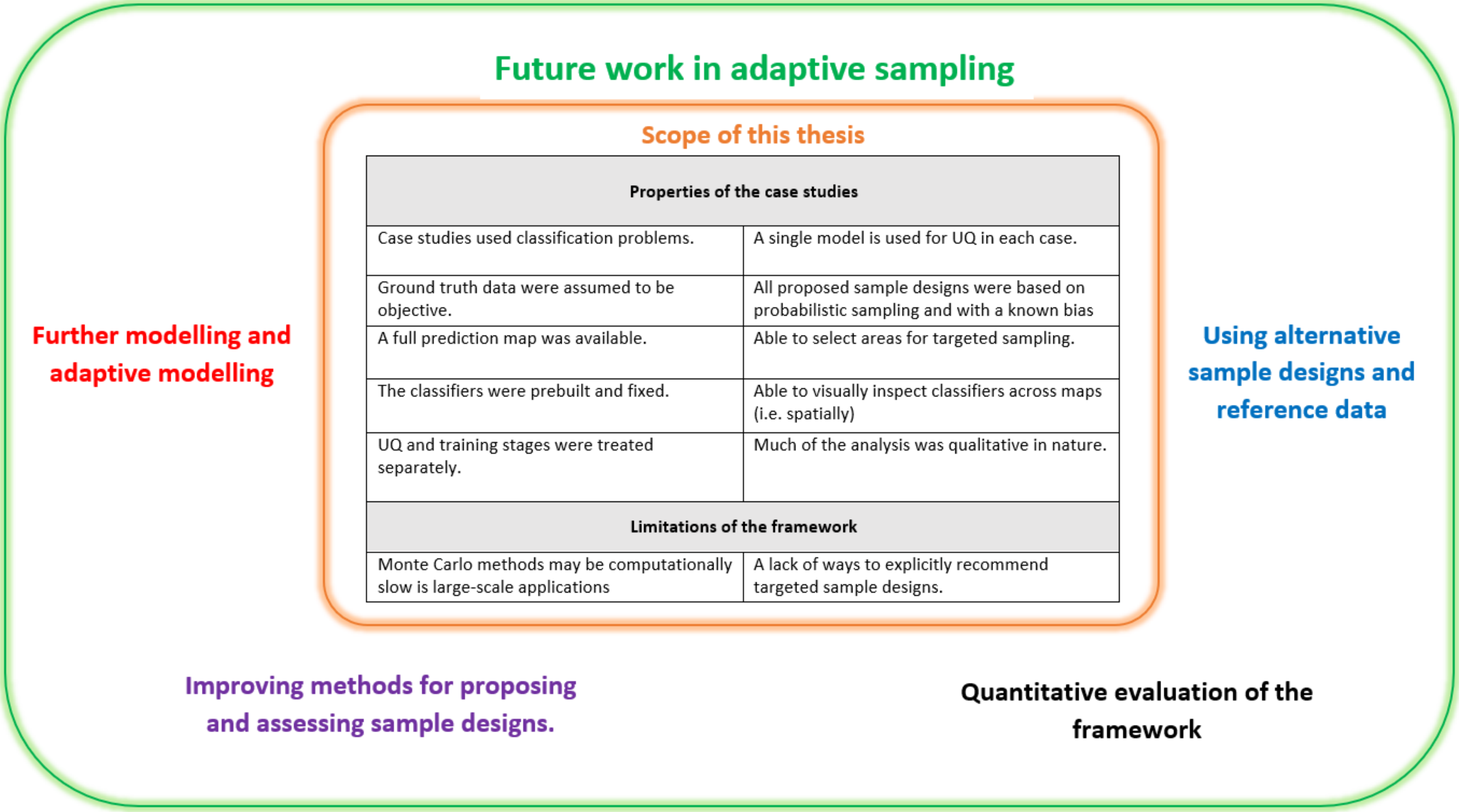
An illustration of the benefits of using Bayesian inference in adaptive sampling.

Whilst not strictly necessary for the framework, this thesis has illustrated that using Bayesian inference brings many benefits in the context of adaptive sampling. This is because many adaptive sampling methods are easier to apply under Bayesian inference as it is naturally better suited to sequential sampling and propagating uncertainty with Monte Carlo methods.

This result is important as differences between Bayesian and frequentist inference can seem trivial when reference samples are large, collected under a single phase and produce numerically similar outputs. However, the advantages of Bayesian inference become much more apparent when reference sampling is limited by design restrictions and using adaptive sampling. Ultimately these advantages should motivate any transition from frequentist to Bayesian inference when looking at adaptive sampling strategies.

7.3 Future work

This section documents important areas of future work under four headings: further modelling and adaptive modelling (7.3.1), using alternative sample designs and reference data (7.3.2), improving methods for proposing and assessing sample designs (7.3.3), and quantitative evaluation of the framework (7.3.4). A summary of the scope of this thesis and how they relate to these topics for further exploration are summarised in Figure 7.1.



Future work in adaptive sampling

Scope of this thesis

Properties of the case studies	
Case studies used classification problems.	A single model is used for UQ in each case.
Ground truth data were assumed to be objective.	All proposed sample designs were based on probabilistic sampling and with a known bias
A full prediction map was available.	Able to select areas for targeted sampling.
The classifiers were prebuilt and fixed.	Able to visually inspect classifiers across maps (i.e. spatially)
UQ and training stages were treated separately.	Much of the analysis was qualitative in nature.
Limitations of the framework	
Monte Carlo methods may be computationally slow is large-scale applications	A lack of ways to explicitly recommend targeted sample designs.

Further modelling and adaptive modelling

Using alternative sample designs and reference data

Improving methods for proposing and assessing sample designs.

Quantitative evaluation of the framework

Figure 7.1. *A summary of the scope of this thesis (orange inner ring) and topics for further exploration are to be discussed in Section 7.3 (outer green ring). The inner ring provides a summary of the thesis by listing the properties of the case studies and the limitations of the framework.*

7.3.1 Further modelling and adaptive modelling

Throughout the case studies in this thesis, there were a number of assumptions or restrictions placed on the modelling used to quantify uncertainty. Some examples of these assumptions and restrictions include using classification problems; using separate data sets for training the classifiers and UQ; assuming no spatial correlation for noise components in models; and knowing the bias in the initial sample designs. An obvious avenue for future work could consider more case studies where some of these assumptions or restrictions differ and improve upon the framework when necessary.

However, simply applying the framework to new scenarios with different model structures may not be enough though, as one still needs to account for the possibility that there may be many plausible approaches for UQ in the adaptive sampling process. Much in the same way that adaptive sampling involves experimenting with different designs under uncertainty and balancing different trade-offs, the choice of models and method used to quantify uncertainty is also subject to uncertainty which may be updated as further data between sampling iterations arrives. Hence, future work may be better focused on the idea of combining adaptive sampling with adaptive modelling, where the models and UQ methods can also change between iterations.

One option would be to consider a shortlist of models and see where they agree and where they disagree across a mapping. In this situation, areas with a large disagreement between models would be an indication that future sampling should target these areas. Furthermore, some models may be added and removed from the shortlist as more reference data becomes available. The generalisability of Monte Carlo methods would help in these situations, as the core methodology is the same across models. An alternative approach would combine shortlisted models into one model through an ensemble approach. For example, one could consider a weighted average of multiple models. Bayesian inference is naturally suited to this, as the weights can easily be included as another model parameter and hence considered in posterior distributions [378], which could act as a more automated (and less abrupt) way of adding and removing or adding models as more reference data are collected.

7.3.2 Using alternative sample designs and reference data

In addition to the modelling and UQ assumptions, there were also assumptions within the sample designs and reference data used throughout the case studies. Three common properties were that (i) all designs were based on probabilistic sampling, (ii) sample designs could explicitly target areas in the mapped area and (iii) the reference data were assumed to be objective and fully deterministic. Since a large part of the adaptive sampling framework is focused on a generalisable approach, a natural area of future work would be to investigate how an adaptive sampling framework could make use of alternative sample designs and types of reference data. In particular, there are three noteworthy avenues for future work in adaptive sampling under alternative designs and reference data: making use of noisy reference data, making use of non-probabilistic sampling, and using adaptive sampling designs when explicit targeting is not available.

Making use of noisy reference data

In this thesis, the reference data were assumed to be objectively correct. Even in the domain of land cover mappings though, this assumption may not hold as uncertainty in reference data can arise from sources such as measurement errors in sensors, disagreements between expert assessments, or when relying on reference data that is not a ground-truth assessment due to practical restrictions (e.g. using aerial photography in place of physical visitations for inaccessible areas) [41], [379], [380], [381]. Providing that one can account for noise in reference data in UQ though, the framework should be able to handle these as it would involve the core processes with some additional sources of uncertainty under different model structures.

Whilst one can account for noisy reference data with the framework in its current form, future work can still look towards how to best balance different levels of noise against sampling convenience when managing uncertainty. Typically, this would involve answering questions such as "*Is it better to get more reference data with a moderate amount of noise or less data with a higher quality assurance?*". From the perspective of the adaptive sampling framework, one may well imagine a situation where cheaper methods of reference sampling may be better at reducing ontological and epistemic uncertainty at the cost of more aleatoric uncertainty due to additional noise. For land cover mappings specifically, such dynamics become more relevant as using reference data from crowdsourcing becomes more popular [382], [383], [384].

Making use of non-probabilistic sampling

The sample designs in the case studies were all based on reducing uncertainty in parameter values under different forms of probabilistic sampling as this was the only viable means of reducing uncertainty due to the simplistic nature of the models and methods in UQ. As one looks towards more advanced modelling structures though, future work could consider using non-probabilistic designs in an adaptive sampling framework, which are generally easier and cheaper to apply.

In particular, model structures where some of the predictors themselves are based on models (i.e. model chains) and models that take into account autocorrelations (e.g. spatial or temporal) are likely to benefit from adaptive sampling under non-probabilistic designs. In the case of model chaining, uncertainty can be reduced using direct measurements for the predictors instead of relying on the modelled values. For models with autocorrelation, having reference data from nearby instances (either spatially or temporally) can be enough to reduce uncertainty. For both examples, there is no requirement for reference data to be collected under probabilistic sampling.

Looking at the adaptive sampling framework, there should not be a problem using the methods under non-probabilistic sample designs, as the principal idea of using Bayesian inference with Monte Carlo methods to experiment with sample designs does not require probabilistic sample designs. However, there are still likely to be interesting questions about how the framework can be used to strike an optimal balance between probabilistic and non-probabilistic sampling under hybrid approaches.

Using adaptive sampling when explicit targeting is not available.

Another feature common to the case studies was the ability to explicitly target areas in sample designs. In the Lagos case study, one could target areas by changing the sample sizes across the strata. In the woodland case study, targeted sampling was achieved by changing the inclusion probabilities across the predictive feature space. This luxury may not hold when looking to manage uncertainty under machine learning classifiers such as insurance fraud [385], [386], [387], early fault detection [388], [389], [390], or hiring processes [391], [392] though.

In these examples, one cannot target different subsets of the population in the same way they could in the land cover case studies as it may be impractical to revisit settled insurance claims, know if a fault would have been detected by a manual inspection at the time, or expect a previously rejected applicant to partake in a recruitment process for a position that may longer exist. In short, with the land cover mapping case studies, it is much easier to apply statements of the form “go collect reference data from these areas” when compared to other applications.

Hence future work could investigate how the adaptive sampling framework can be applied when one cannot explicitly target areas of a population. This could involve prioritising the kinds of instances one collects reference data from as opposed to explicitly targeting areas. Essentially, this change in approach would take the previous “go collect reference data from these areas” statements to something more like “when an instance from this area comes along, make sure you collect some reference data for it”.

7.3.3 Improving methods for proposing and assessing sample designs

A recurring theme throughout the thesis is that a large component of adaptive sampling involves exploring and experimenting with different sample designs. This explorative approach to sample design works best when one has the means to propose suitable sample designs and the analytical tools to assess them before committing to the designs. Overall, the case studies have illustrated how the methods in the framework can be used to meet these goals. That said, there are still opportunities in future work to improve how sample designs are proposed and assessed. Three opportunities worth considering here are (i) overcoming computational challenges in Monte Carlo methods and Bayesian inference, (ii) improving the analytical tools in design assessment, and (iii), developing ways to automatically recommend good sample designs.

Overcoming the computational challenges in Monte Carlo methods and Bayesian inference

One of the key strengths of the adaptive framework comes from the use of Bayesian inference and Monte Carlo methods which in turn makes the framework highly generalisable. However, the combination of Bayesian inference and Monte Carlo methods can be computationally expensive in large-scale applications, which in turn negatively impacts the ability to experiment with sample design and modelling possibilities. In other words, it becomes difficult to explore and experiment with different options when the assessment of each design takes a long time to compute.

Consequently, anything that can help reduce the time it takes to assess sample designs (and potentially model choices) when using Monte Carlo methods will indirectly help improve the framework. There are several potential options here. The use of cloud computing has seen success in recent mapping applications at a more general level [393], [394], [395], [396], so folding the adaptive sampling framework into this is one option. Other options may look toward increasing computational power with neuromorphic computing [397], [398] or developing methods that make sampling from posterior distributions more efficient [399], [400], [401]. It is important to note though that these options are topics that are arguably areas of research in their own right. Hence, it may be best to view the computational challenges in Monte Carlo methods in Bayesian inference as related to, but separate from, the adaptive sampling framework.

Improving the analytical tools in design assessment

Under the case studies visualising the results in the design assessment phase was relatively straightforward. The first part of the Lagos case study focused on a point estimate (i.e. the total urbanised area) which made comparing the likely effects of further sampling under different designs possible through box plots. For the second part of the Lagos case study and the England case study, results could be viewed across the predictive feature space with 2 and 3-dimensional plots respectively and spatially across the mapped areas. This ability to visualise the assessments helped a lot when experimenting with different sample designs. However, the ability to view results across low-dimensional feature spaces or spatially may not hold across all applications.

With this in mind, future work could focus on improving the analytical tools for assessing different sample designs. This could involve developing visualisation tools and metrics that can handle models in higher dimensional settings.

Developing ways to automatically recommend good sample designs.

While exploring and experimenting with different sample designs can offer many advantages in adaptive sampling, finding suitable sample designs to begin with may not always be so easy. Given the previously discussed issues of computational costs and the potential difficulties in visualising results in higher dimensional settings, future work may consider ways of automatically generating good sample designs that can act as reasonable starting points for experimentation. These methods could be motivated by optimisation problems. For example, one may consider investigating how Bayesian optimisation techniques could help in generating sample designs when uncertainty is measured using entropy [402], [403]. Equally, one could take a less formal approach to make use of machine learning techniques (e.g. detecting clusters of high epistemic uncertainty for targeted sampling). Either way, one can see natural links between improving the analytical tools in design assessment and developing ways to automatically recommend good sample designs.

7.3.4 Quantitative evaluation of the framework

Much of the analysis in this thesis was deliberately qualitative. This was necessary because the focus of the thesis was on developing and evaluating a newly established framework that aimed to deal with the problem of managing uncertainty in machine learning at a more general level. In many ways, the framework was not ready for quantitative methods of analysis, as it still had to undergo refinement and be evaluated at a level that could not be easily tested using quantitative analysis. For example, there is no clear quantitative measure that captures how using a model's predictors as a basis for targeted sampling can greatly simplify the process of quantifying uncertainty under iterative sampling.

However, with the adaptive framework (and its qualitative benefits) more firmly established, a natural area for future work would be to evaluate the framework quantitatively using various performance metrics. Examples of performance metrics here could include computational costs and running times; measures that capture the precision of estimates, total monetary savings etc. Ideally, any quantitative evaluation would be applied across a representative set of modelling scenarios, as this would align with the goal of creating a generalisable framework.

7.4 Concluding remarks

As techniques in machine learning are further developed and employed across more applications, the need to efficiently manage uncertainty under these methods only grows. Whilst there is a lot of work on managing uncertainty under more traditional modelling environments, bringing these methods into machine learning applications comes with many challenges as machine learning techniques often lack transparency and require large volumes of data to train.

To overcome these challenges, this thesis has proposed an adaptive sampling framework that is agnostic to the choice of machine learning classification and quasi-agnostic to the method of uncertainty quantification (i.e. the framework does not depend on specific forms of uncertainty quantification, but Bayesian inference comes highly recommended).

The motivation for this was that by building sample designs iteratively and keeping the framework agnostic to the choice of machine learning techniques and the method of uncertainty quantification, one could offer a more consistent and generalisable way of managing uncertainty.

Overall, the results from this thesis are promising. Whilst the thesis has focused on case studies in land cover mapping applications to evaluate the framework, the results and principles easily transfer to other applications. Consequently, the framework should benefit a wide variety of situations where there is a need to efficiently manage uncertainty in machine learning techniques due to design restrictions when sampling reference data.

References.

[1] A. Basit, M. Zafar, X. Liu, A. R. Javed, Z. Jalil, and K. Kifayat, ‘A comprehensive survey of AI-enabled phishing attacks detection techniques’, *Telecommunication Systems*. 2021. doi: 10.1007/s11235-020-00733-2.

[2] W. Ma, D. Tran, and D. Sharma, ‘A novel spam email detection system based on negative selection’, in *ICCIT 2009 - 4th International Conference on Computer Sciences and Convergence Information Technology*, 2009. doi: 10.1109/ICCIT.2009.58.

[3] I. Idris and A. Selamat, ‘Improved email spam detection model with negative selection algorithm and particle swarm optimization’, *Applied Soft Computing Journal*, 2014, doi: 10.1016/j.asoc.2014.05.002.

[4] O. Russakovsky *et al.*, ‘ImageNet Large Scale Visual Recognition Challenge’, *Int J Comput Vis*, 2015, doi: 10.1007/s11263-015-0816-y.

[5] B. Vijaya Kumar and K. Bhavya, ‘Dog breed identification with fine tuning of pre-trained models’, *International Journal of Recent Technology and Engineering*, 2019, doi: 10.35940/ijrte.B1464.0982S1119.

[6] Z. Ráduly, C. Sulyok, Z. Vadászi, and A. Zölde, ‘Dog Breed Identification Using Deep Learning’, in *SISY 2018 - IEEE 16th International Symposium on Intelligent Systems and Informatics, Proceedings*, 2018. doi: 10.1109/SISY.2018.8524715.

[7] M. Fatima and M. Pasha, ‘Survey of Machine Learning Algorithms for Disease Diagnostic’, *Journal of Intelligent Learning Systems and Applications*, 2017, doi: 10.4236/jilsa.2017.91001.

[8] R. Alizadehsani *et al.*, ‘Machine learning-based coronary artery disease diagnosis: A comprehensive review’, *Computers in Biology and Medicine*. 2019. doi: 10.1016/j.compbiomed.2019.103346.

[9] D. Jain and V. Singh, ‘Feature selection and classification systems for chronic disease prediction: A review’, *Egyptian Informatics Journal*. 2018. doi: 10.1016/j.eij.2018.03.002.

[10] S. Luo, X. Li, and J. Li, ‘Automatic Alzheimer’s Disease Recognition from MRI Data Using Deep Learning Method’, *Journal of Applied Mathematics and Physics*, 2017, doi: 10.4236/jamp.2017.59159.

[11] S. Benson Edwin Raj and A. Annie Portia, ‘Analysis on credit card fraud detection methods’, in *2011 International Conference on Computer, Communication and Electrical Technology, ICCCET 2011*, 2011. doi: 10.1109/ICCCET.2011.5762457.

[12] D. Varmedja, M. Karanovic, S. Sladojevic, M. Arsenovic, and A. Anderla, ‘Credit Card Fraud Detection - Machine Learning methods’, in *2019 18th International Symposium INFOTEH-JAHORINA, INFOTEH 2019 - Proceedings*, 2019. doi: 10.1109/INFOTEH.2019.8717766.

[13] J. O. Awoyemi, A. O. Adetunmbi, and S. A. Oluwadare, ‘Credit card fraud detection using machine learning techniques: A comparative analysis’, in *Proceedings of the IEEE International Conference on Computing, Networking and Informatics, ICCNI 2017*, 2017. doi: 10.1109/ICCNI.2017.8123782.

[14] I. Sakharova, ‘Payment card fraud: Challenges and solutions’, in *ISI 2012 - 2012 IEEE International Conference on Intelligence and Security Informatics: Cyberspace, Border, and Immigration Securities*, 2012. doi: 10.1109/ISI.2012.6284315.

[15] A. L. Fradkov, ‘Early History of Machine Learning’, *IFAC-PapersOnLine*, vol. 53, no. 2, pp. 1385–1390, 2020, doi: https://doi.org/10.1016/j.ifacol.2020.12.1888.

[16] N. Abramson, D. Braverman, and G. Sebestyen, ‘Pattern recognition and machine learning’, *IEEE Trans Inf Theory*, 1963, doi: 10.1109/TIT.1963.1057854.

[17] C. A. Rosen, ‘Pattern classification by adaptive machines’, *Science (1979)*, 1967, doi: 10.1126/science.156.3771.38.

[18] A.~L.~Samuel, ‘Some Studies in Machine Learning Using the Game of Checkers’, *IBM J Res Dev*, 1959.

[19] A. L. Samuel, ‘Some studies in machine learning using the game of checkers. II-Recent progress’, *Annual Review in Automatic Programming*. 1969. doi: 10.1016/0066-4138(69)90004-4.

[20] I. Kononenko, ‘Machine learning for medical diagnosis: History, state of the art and perspective’, *Artif Intell Med*, 2001, doi: 10.1016/S0933-3657(01)00077-X.

[21] B. J. Erickson, P. Korfiatis, Z. Akkus, and T. L. Kline, ‘Machine learning for medical imaging’, *Radiographics*, 2017, doi: 10.1148/rg.2017160130.

[22] A. Ozcift and A. Gulten, ‘Classifier ensemble construction with rotation forest to improve medical diagnosis performance of machine learning algorithms’, *Comput Methods Programs Biomed*, 2011, doi: 10.1016/j.cmpb.2011.03.018.

[23] M. Pak and S. Kim, ‘A review of deep learning in image recognition’, in *Proceedings of the 2017 4th International Conference on Computer Applications and Information Processing Technology, CAIPT 2017*, 2018. doi: 10.1109/CAIPT.2017.8320684.

[24] R. Dash and P. K. Dash, ‘A hybrid stock trading framework integrating technical analysis with machine learning techniques’, *Journal of Finance and Data Science*, 2016, doi: 10.1016/j.jfds.2016.03.002.

[25] J. Patel, S. Shah, P. Thakkar, and K. Kotecha, ‘Predicting stock and stock price index movement using Trend Deterministic Data Preparation and machine learning techniques’, *Expert Syst Appl*, 2015, doi: 10.1016/j.eswa.2014.07.040.

[26] D. Lv, S. Yuan, M. Li, and Y. Xiang, ‘An Empirical Study of Machine Learning Algorithms for Stock Daily Trading Strategy’, *Math Probl Eng*, 2019, doi: 10.1155/2019/7816154.

[27] A. Saranya and R. Anandan, ‘Stock market prediction using machine learning algorithms’, *International Journal of Recent Technology and Engineering*, 2019, doi: 10.35940/ijrte.B1052.0782S419.

[28] P. A. O'Gorman and J. G. Dwyer, 'Using Machine Learning to Parameterize Moist Convection: Potential for Modeling of Climate, Climate Change, and Extreme Events', *J Adv Model Earth Syst*, 2018, doi: 10.1029/2018MS001351.

[29] V. M. Krasnopolsky and M. S. Fox-Rabinovitz, 'Complex hybrid models combining deterministic and machine learning components for numerical climate modeling and weather prediction', *Neural Networks*, 2006, doi: 10.1016/j.neunet.2006.01.002.

[30] V. M. Krasnopolsky, M. S. Fox-Rabinovitz, and D. V. Chalikov, 'New approach to calculation of atmospheric model physics: Accurate and fast neural network emulation of longwave radiation in a climate model', *Mon Weather Rev*, 2005, doi: 10.1175/MWR2923.1.

[31] Tom Mitchell, 'Machine Learning textbook', McGraw Hill.

[32] K. Beven, 'Facets of uncertainty: Epistemic uncertainty, non-stationarity, likelihood, hypothesis testing, and communication', *Hydrological Sciences Journal*, 2016, doi: 10.1080/02626667.2015.1031761.

[33] S. Talukdar *et al.*, 'Land-use land-cover classification by machine learning classifiers for satellite observations-A review', *Remote Sensing*. 2020. doi: 10.3390/rs12071135.

[34] H. Keshtkar, W. Voigt, and E. Alizadeh, 'Land-cover classification and analysis of change using machine-learning classifiers and multi-temporal remote sensing imagery', *Arabian Journal of Geosciences*, 2017, doi: 10.1007/s12517-017-2899-y.

[35] C. R. Fichera, G. Modica, and M. Pollino, 'Land Cover classification and change-detection analysis using multi-temporal remote sensed imagery and landscape metrics', *Eur J Remote Sens*, 2012, doi: 10.5721/EuJRS20124501.

[36] C. M. Viana, I. Girão, and J. Rocha, 'Long-term satellite image time-series for land use/land cover change detection using refined open source data in a rural region', *Remote Sens (Basel)*, 2019, doi: 10.3390/rs11091104.

[37] N. Horning, E. Fleishman, P. J. Ersts, F. A. Fogarty, and M. Wohlfeil Zillig, 'Mapping of land cover with open-source software and ultra-high-resolution imagery acquired with unmanned aerial vehicles', *Remote Sens Ecol Conserv*, 2020, doi: 10.1002/rse2.144.

[38] D. Ienco, R. Interdonato, R. Gaetano, and D. Ho Tong Minh, 'Combining Sentinel-1 and Sentinel-2 Satellite Image Time Series for land cover mapping via a multi-source deep learning architecture', *ISPRS Journal of Photogrammetry and Remote Sensing*, 2019, doi: 10.1016/j.isprsjprs.2019.09.016.

[39] D. Phiri, M. Simwanda, S. Salekin, V. R. Nyirenda, Y. Murayama, and M. Ranagalage, 'Sentinel-2 data for land cover/use mapping: A review', *Remote Sensing*. 2020. doi: 10.3390/rs12142291.

[40] P. Olofsson, G. M. Foody, M. Herold, S. V. Stehman, C. E. Woodcock, and M. A. Wulder, 'Good practices for estimating area and assessing accuracy of land change', *Remote Sens Environ*, vol. 148, pp. 42–57, 2014, doi: 10.1016/j.rse.2014.02.015.

[41] G. M. Foody, 'Status of land cover classification accuracy assessment', *Remote Sensing of Environment*. 2002. doi: 10.1016/S0034-4257(01)00295-4.

[42] M. B. Lyons, D. A. Keith, S. R. Phinn, T. J. Mason, and J. Elith, ‘A comparison of resampling methods for remote sensing classification and accuracy assessment’, *Remote Sens Environ*, 2018, doi: 10.1016/j.rse.2018.02.026.

[43] G. Ståhl *et al.*, ‘Use of models in large-area forest surveys: comparing model-assisted, model-based and hybrid estimation’, *Forest Ecosystems*. 2016. doi: 10.1186/s40663-016-0064-9.

[44] J. Pickering *et al.*, ‘Quantifying the trade-off between cost and precision in estimating area of forest loss and degradation using probability sampling in Guyana’, *Remote Sens Environ*, 2019, doi: 10.1016/j.rse.2018.11.018.

[45] U.S. Geological Survey, ‘LC08 L1TP 191055 20160210 20170330 01 T1’. Accessed: Jul. 30, 2020. [Online]. Available: https://earthexplorer.usgs.gov/scene/metadata/full/5e83d0b656b77cf3/LC81910552016041LGN01/

[46] L. Breiman, ‘Random forests’, *Mach. Learn.*, 2001, doi: 10.1023/A:1010933404324.

[47] R. J. Little, ‘To model or not to model? Competing modes of inference for finite population sampling’, *Journal of the American Statistical Association*. 2004. doi: 10.1198/016214504000000467.

[48] M. H. Hansen, W. G. Madow, and B. J. Tepping, ‘An evaluation of model-dependent and probability-sampling inferences in sample surveys’, *J Am Stat Assoc*, 1983, doi: 10.1080/01621459.1983.10477018.

[49] C.-E. Särndal, I. Thomsen, J. M. Hoem, D. V Lindley, O. Barndorff-Nielsen, and T. Dalenius, ‘Design-Based and Model-Based Inference in Survey Sampling [with Discussion and Reply]’, *Scandinavian Journal of Statistics*, 1978.

[50] L. C. Zhang, ‘On valid descriptive inference from non-probability sample’, *Stat Theory Relat Fields*, 2019, doi: 10.1080/24754269.2019.1666241.

[51] Y. Shi, C. J. Cameron, and D. D. Heckathorn, ‘Model-Based and Design-Based Inference: Reducing Bias Due to Differential Recruitment in Respondent-Driven Sampling’, *Sociol Methods Res*, 2019, doi: 10.1177/0049124116672682.

[52] T. G. Gregoire, ‘Design-based and model-based inference in survey sampling: Appreciating the difference’, *Canadian Journal of Forest Research*. 1998. doi: 10.1139/x98-166.

[53] H. T. Schreuder, T. G. Gregoire, and J. P. Weyer, ‘For what applications can probability and non-probability sampling be used?’, *Environ Monit Assess*, 2001, doi: 10.1023/A:1006316418865.

[54] D. J. Brus and J. J. De Gruijter, ‘Random sampling or geostatistical modelling? Choosing between design-based and model-based sampling strategies for soil (with Discussion)’, *Geoderma*. 1997. doi: 10.1016/S0016-7061(97)00072-4.

[55] S. Geuna, ‘Appreciating the difference between design-based and model-based sampling strategies in quantitative morphology of the nervous system’, *Journal of Comparative Neurology*. 2000. doi: 10.1002/1096-9861(20001120)427:3<333::AID-CNE1>3.0.CO;2-T.

[56] S. V. Stehman, ‘Practical implications of design-based sampling inference for thematic map accuracy assessment’, *Remote Sens Environ,* 2000, doi: 10.1016/S0034-4257(99)00090-5.

[57] S. F. Crone, S. Lessmann, and R. Stahlbock, ‘The impact of preprocessing on data mining: An evaluation of classifier sensitivity in direct marketing’, *Eur J Oper Res*, 2006, doi: 10.1016/j.ejor.2005.07.023.

[58] S. K. Sterba, ‘Alternative model-based and design-based frameworks for inference from samples to populations: From polarization to integration’, *Multivariate Behav Res*, 2009, doi: 10.1080/00273170903333574.

[59] C.-E. Särndal, B. Swensson, and J. Wretman, *Model-Assisted Survey Sampling*. Springer Science & Business Media, 2003.

[60] J. M. Hunter *et al.*, ‘Framework for developing hybrid process-driven, artificial neural network and regression models for salinity prediction in river systems’, *Hydrol. Earth Syst. Sci*, vol. 22, pp. 2987–3006, 2018, doi: 10.5194/hess-22-2987-2018.

[61] M. Hesamia, R. Naderia, M. Yoosefzadeh-Najafabadia, and Mostafa Rahmatib, ‘Data-Driven Modeling in Plant Tissue Culture’, *J. Appl. Environ. Biol. Sci.*, vol. 7, no. 8, pp. 37–44, 2017.

[62] F. J. Montáns, F. Chinesta, R. Gómez-Bombarelli, and J. N. Kutz, ‘Data-driven modeling and learning in science and engineering’, *Comptes Rendus - Mecanique*. 2019. doi: 10.1016/j.crme.2019.11.009.

[63] Z. Ge, ‘Review on data-driven modeling and monitoring for plant-wide industrial processes’, *Chemometrics and Intelligent Laboratory Systems*. 2017. doi: 10.1016/j.chemolab.2017.09.021.

[64] D. B. Searls, ‘Data integration: Challenges for drug discovery’, *Nature Reviews Drug Discovery*. 2005. doi: 10.1038/nrd1608.

[65] B. Lu and B. R. Upadhyaya, ‘Monitoring and fault diagnosis of the steam generator system of a nuclear power plant using data-driven modeling and residual space analysis’, *Ann Nucl Energy*, 2005, doi: 10.1016/j.anucene.2005.02.003.

[66] C. Rudin, ‘Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead’, *Nature Machine Intelligence*. 2019. doi: 10.1038/s42256-019-0048-x.

[67] F. K. Dosilovic, M. Brcic, and N. Hlupic, ‘Explainable artificial intelligence: A survey’, in *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2018 - Proceedings*, 2018. doi: 10.23919/MIPRO.2018.8400040.

[68] E.-J. Wagenmakers, M. Lee, T. Lodewyckx, and G. J. Iverson, ‘Bayesian Versus Frequentist Inference’, in *Bayesian Evaluation of Informative Hypotheses*, 2008. doi: 10.1007/978-0-387-09612-4_9.

[69] M. J. Bayarri and J. O. Berger, ‘The interplay of Bayesian and frequentist analysis’, *Statistical Science*, 2004, doi: 10.1214/088342304000000116.

[70] J. Neyman, ‘Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability’, *Philosophical Transactions of the Royal Society of London. Series A,*

Mathematical and Physical Sciences, vol. 236, no. 767, p. DOI:https://doi.org/10.1098/rsta.1937.0005 Publish, 1937.

[71] D. Berrar and J. A. Lozano, 'Significance tests or confidence intervals: Which are preferable for the comparison of classifiers?', *Journal of Experimental and Theoretical Artificial Intelligence*, 2013, doi: 10.1080/0952813X.2012.680252.

[72] K. Dunnigan, 'Confidence Interval Calculation for Binomial Proportions', *Mwsug*, p. 12, 2008.

[73] B. Efron and R. Tibshirani, 'Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy', *Statistical Science*, 1986, doi: 10.1214/ss/1177013815.

[74] C. Cortes and M. Mohri, 'Confidence intervals for the area under the ROC Curve', in *Advances in Neural Information Processing Systems*, 2005.

[75] E. Ledell, M. Petersen, and M. Van Der Laan, 'Computationally efficient confidence intervals for cross-validated area under the ROC curve estimates', *Electron J Stat*, 2015, doi: 10.1214/15-EJS1035.

[76] R. K. Steinhorst and R. H. Myers, 'Classical and Modern Regression With Applications.', *J Am Stat Assoc*, 1988, doi: 10.2307/2288958.

[77] P. Morris, C.-E. Sarndal, B. Swensson, and J. Wretman, 'Model Assisted Survey Sampling', *The Mathematical Gazette*, 1993, doi: 10.2307/3619754.

[78] K. R. Koch, *Introduction to bayesian statistics*. 2007. doi: 10.1007/978-3-540-72726-2.

[79] A. Etz and J. Vandekerckhove, 'Introduction to Bayesian Inference for Psychology', *Psychon Bull Rev*, 2018, doi: 10.3758/s13423-017-1262-3.

[80] R. E. Kass and L. Wasserman, 'The selection of prior distributions by formal rules', *J Am Stat Assoc*, 1996, doi: 10.1080/01621459.1996.10477003.

[81] G. E. P. Box and G. C. Tiao, *Bayesian Inference in Statistical Analysis*. 1992. doi: 10.1002/9781118033197.

[82] A. M. Gressner and O. A. Gressner, 'A Compendium of Conjugate Priors', *Lexikon der Medizinischen Laboratoriumsdiagnostik*, 2018, doi: 10.1007/978-3-662-49054-9_3064-1.

[83] P. Diaconis and D. Ylvisaker, 'Conjugate Priors for Exponential Families', *The Annals of Statistics*, 1979, doi: 10.1214/aos/1176344611.

[84] S. R. Dalal and W. J. Hall, 'Approximating Priors by Mixtures of Natural Conjugate Priors', *Journal of the Royal Statistical Society: Series B (Methodological)*, 1983, doi: 10.1111/j.2517-6161.1983.tb01251.x.

[85] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian data analysis, third edition*. 2013.

[86] L. Hespanhol, C. S. Vallio, L. M. Costa, and B. T. Saragiotto, 'Understanding and interpreting confidence and credible intervals around effect estimates', *Braz J Phys Ther*, vol. 23, no. 4, pp. 290–301, Jul. 2019, doi: 10.1016/J.BJPT.2018.12.006.

[87] F. J. Samaniego, 'A comparison of the Bayesian and frequentist approaches to estimation', *Media*, 2010.

[88] N. Balakrishnan, E. Cramer, and G. Iliopoulos, ‘On the method of pivoting the CDF for exact confidence intervals with illustration for exponential mean under life-test with time constraints’, *Stat Probab Lett*, 2014, doi: 10.1016/j.spl.2014.02.022.

[89] C. J. Clopper and Pearson, ‘The use of confidence or fiducial limits illustrated in the case of the binomial’, *Biometrika*, pp. 404–413, 1934.

[90] M. Thulin, ‘The cost of using exact confidence intervals for a binomial proportion’, *Electron J Stat*, 2014, doi: 10.1214/14-EJS909.

[91] A. Agresti and B. A. Coull, ‘Approximate Is Better than " Exact " for Interval Estimation of Binomial Proportions Published by : Taylor & Francis , Ltd . on behalf of the American Statistical Association Stable URL : http://www.jstor.org/stable/2685469 Approximate is Better than " Ex’, *Am Stat*, vol. 52, no. 2, pp. 119–126, 1998.

[92] E. Cameron, ‘On the estimation of confidence intervals for binomial population proportions in astronomy: The simplicity and superiority of the Bayesian approach’, *Publications of the Astronomical Society of Australia*, 2011, doi: 10.1071/AS10046.

[93] S. G. Kwak and J. H. Kim, ‘Central limit theorem: The cornerstone of modern statistics’, *Korean J Anesthesiol*, 2017, doi: 10.4097/kjae.2017.70.2.144.

[94] P. J. Bickel and D. A. Freedman, ‘Some Asymptotic Theory for the Bootstrap’, *The Annals of Statistics*, 1981, doi: 10.1214/aos/1176345637.

[95] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*. Boca Raton, FL: CHAPMAN & HALLICR, 1993.

[96] J. L. Peugh, ‘A practical guide to multilevel modeling’, *J Sch Psychol*, 2010, doi: 10.1016/j.jsp.2009.09.002.

[97] S. G. Self and K. Y. Liang, ‘Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions’, *J Am Stat Assoc*, 1987, doi: 10.1080/01621459.1987.10478472.

[98] M. H. C. Lai, ‘Bootstrap Confidence Intervals for Multilevel Standardized Effect Size’, *Multivariate Behav Res*, 2021, doi: 10.1080/00273171.2020.1746902.

[99] S. J. Pocock, ‘Group sequential methods in the design and analysis of clinical trials’, *Biometrika*, 1977, doi: 10.1093/biomet/64.2.191.

[100] Y. Cheng and Y. Shen, ‘Estimation of a parameter and its exact confidence interval following sequential sample size reestimation trials’, *Biometrics*, 2004, doi: 10.1111/j.0006-341X.2004.00246.x.

[101] W. Lehmacher and G. Wassmer, ‘Adaptive sample size calculations in group sequential trials’, *Biometrics*, 1999, doi: 10.1111/j.0006-341X.1999.01286.x.

[102] C. Jennison and B. W. Turnbull, ‘Meta-analyses and adaptive group sequential designs in the clinical development process’, *Journal of Biopharmaceutical Statistics*. 2005. doi: 10.1081/BIP-200062273.

[103] L. E. Bothwell, J. Avorn, N. F. Khan, and A. S. Kesselheim, ‘Adaptive design clinical trials: A review of the literature and ClinicalTrials.gov’, *BMJ Open*, 2018, doi: 10.1136/bmjopen-2017-018320.

[104] J. S. Pontius and M. C. Christman, 'BOOTSTRAP CONFIDENCE INTERVALS FROM ADAPTIVE SAMPLING OF AN INSECT POPULATION', *Conference on Applied Statistics in Agriculture*, Apr. 1997, doi: 10.4148/2475-7772.1307.

[105] J. Geweke, 'Bayesian Inference in Econometric Models Using Monte Carlo Integration', *Econometrica*, 1989, doi: 10.2307/1913710.

[106] C. J. Geyer, 'Introduction to Markov Chain Monte Carlo', *Handbook of Markov Chain Monte Carlo*, 2011.

[107] D. van Ravenzwaaij, P. Cassey, and S. D. Brown, 'A simple introduction to Markov Chain Monte–Carlo sampling', *Psychon Bull Rev*, 2018, doi: 10.3758/s13423-016-1015-8.

[108] D. Stegmueller, 'How many countries for multilevel modeling? A comparison of frequentist and bayesian approaches', *Am J Pol Sci*, 2013, doi: 10.1111/ajps.12001.

[109] A. Benavoli, G. Corani, J. Demšar, and M. Zaffalon, 'Time for a change: A tutorial for comparing multiple classifiers through Bayesian analysis', *Journal of Machine Learning Research*, 2017.

[110] J. Phillipson, G. Blair, and P. Henrys, 'Quantifying Uncertainty for Estimates Derived from Error Matrices in Land Cover Mapping Applications: The Case for a Bayesian Approach', in *IFIP Advances in Information and Communication Technology*, 2020. doi: 10.1007/978-3-030-39815-6_15.

[111] Z. Oravecz, M. Huentelman, and J. Vandekerckhove, 'Sequential bayesian updating for big data', in *Big Data in Cognitive Science*, 2016. doi: 10.4324/9781315413570.

[112] D. V. Lindley, *Bayesian statistics: A review*. Society for industrial and applied mathematics, 1972.

[113] K. S. McConville, G. G. Moisen, and T. S. Frescino, 'A tutorial on model-assisted estimation with application to forest inventory', *Forests*, 2020, doi: 10.3390/f11020244.

[114] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, 'Equation of state calculations by fast computing machines', *J Chem Phys*, vol. 21, no. 6, pp. 1087–1092, 1953, doi: 10.1063/1.1699114.

[115] I. Yildirim, 'Bayesian inference: Metropolis-hastings sampling'. Dept. of Brain and Cognitive Sciences, Univ. of Rochester, Rochester, NY, 2012.

[116] J. S. Rosenthal, 'Optimal proposal distributions and adaptive MCMC', in *Handbook of Markov Chain Monte Carlo*, 2011. doi: 10.1201/b10905-5.

[117] J. A. Christen and C. Foxy, 'A general purpose sampling algorithm for continuous distributions (the t-walk)', *Bayesian Anal*, 2010, doi: 10.1214/10-BA603.

[118] G. O. Roberts, A. Gelman, and W. R. Gilks, 'Weak convergence and optimal scaling of random walk Metropolis algorithms', *Annals of Applied Probability*, 1997, doi: 10.1214/aoap/1034625254.

[119] H. Haario, E. Saksman, and J. Tamminen, 'Adaptive proposal distribution for random walk Metropolis algorithm', *Comput Stat*, 1999, doi: 10.1007/s001800050022.

[120] D. M. Walker, F. J. Pérez-Barbería, and G. Marion, 'Stochastic modelling of ecological processes using hybrid Gibbs samplers', *Ecol Modell*, 2006, doi: 10.1016/j.ecolmodel.2006.04.008.

[121] M. J. Brewer, C. G. G. Aitken, and M. Talbot, 'A comparison of hybrid strategies for Gibbs sampling in mixed graphical models', *Comput Stat Data Anal*, 1996, doi: 10.1016/0167-9473(94)00017-4.

[122] P. Resnik, P. Resnik, E. Hardisty, and E. Hardisty, 'Gibbs Sampling for the Uninitiated', *Umiacs.Umd.Edu*, 2009, doi: 10.1017/CBO9781107415324.004.

[123] A. Kousathanas, C. Leuenberger, J. Helfer, M. Quinodoz, and M. Foll, 'Likelihood-free inference in high-dimensional models', *Genetics*, 2016, doi: 10.1534/genetics.116.187567.

[124] B. M. Turner and T. Van Zandt, 'A tutorial on approximate Bayesian computation', *J Math Psychol*, 2012, doi: 10.1016/j.jmp.2012.02.005.

[125] D. B. Rubin, 'The bayesian bootstrap', *The annals of statistics*, pp. 130–134, 1981.

[126] V. V. Arutyunov, 'Cloud computing: Its history of development, modern state, and future considerations', *Scientific and Technical Information Processing*, 2012, doi: 10.3103/S0147688212030082.

[127] J. Surbiryala and C. Rong, 'Cloud computing: History and overview', in *Proceedings - 2019 3rd IEEE International Conference on Cloud and Fog Computing Technologies and Applications, Cloud Summit 2019*, 2019. doi: 10.1109/CloudSummit47114.2019.00007.

[128] A. S. Acharya, A. Prakash, and A. Nigam, 'Sampling : Why and How of it ? Anita S Acharya , Anupam Prakash , Pikee Saxena ', *Indian Journal of Medical Specialties*, 2013.

[129] G. Sharma, 'Pros and cons of different sampling techniques', *International Journal of Applied Research*, 2017.

[130] I. Etikan, 'Sampling and Sampling Methods', *Biom Biostat Int J*, 2017, doi: 10.15406/bbij.2017.05.00149.

[131] J. F. Etter and T. V. Perneger, 'Snowball sampling by mail: Application to a survey of smokers in the general population', *Int J Epidemiol*, 2000, doi: 10.1093/ije/29.1.43.

[132] R. Miller and J. Brewer, 'Sampling, snowball: accessing hidden and hard-to-reach populations', in *The A-Z of Social Research*, 2016. doi: 10.4135/9780857020024.n94.

[133] D. J. Brus, 'Sampling for digital soil mapping: A tutorial supported by R scripts', *Geoderma*. 2019. doi: 10.1016/j.geoderma.2018.07.036.

[134] R. A. Viscarra Rossel, D. J. Brus, C. Lobsey, Z. Shi, and G. McLachlan, 'Baseline estimates of soil organic carbon by proximal sensing: Comparing design-based, model-assisted and model-based inference', *Geoderma*, 2016, doi: 10.1016/j.geoderma.2015.11.016.

[135] R. E. McRoberts, E. Næsset, and T. Gobakken, 'Estimation for inaccessible and non-sampled forest areas using model-based inference and remotely sensed auxiliary information', *Remote Sens Environ*, 2014, doi: 10.1016/j.rse.2014.08.028.

[136] R. E. McRoberts, ‘Probability- and model-based approaches to inference for proportion forest using satellite imagery as ancillary data’, *Remote Sens Environ*, 2010, doi: 10.1016/j.rse.2009.12.013.

[137] J. M. Brick, ‘Explorations in non-probability sampling using the web’, in *Proceedings of the Conference on beyond traditional survey taking: Adapting to a changing world*, 2014, pp. 1–6.

[138] P. R. Rosenbaum and D. B. Rubin, ‘The central role of the propensity score in observational studies for causal effects’, *Biometrika*, 1983, doi: 10.1093/biomet/70.1.41.

[139] L. Castro-Martín, M. D. M. Rueda, and R. Ferri-García, ‘Estimating general parameters from non-probability surveys using propensity score adjustment’, *Mathematics*, 2020, doi: 10.3390/math8112096.

[140] J. J. Randolph, K. Falbe, A. K. Manuel, and J. L. Balloun, ‘A step-by-step guide to propensity score matching in R’, *Practical Assessment, Research and Evaluation*, 2014.

[141] J. Lee and T. D. Little, ‘A practical guide to propensity score analysis for applied clinical research’, *Behaviour Research and Therapy*, 2017, doi: 10.1016/j.brat.2017.01.005.

[142] K. Hirano, G. W. Imbens, and G. Ridder, ‘Efficient estimation of average treatment effects using the estimated propensity score’, *Econometrica*, 2003, doi: 10.1111/1468-0262.00442.

[143] J. M. Robins, M. Á. Hernán, and B. Brumback, ‘Marginal structural models and causal inference in epidemiology’, *Epidemiology*, 2000, doi: 10.1097/00001648-200009000-00011.

[144] G. Loosveldt and N. Sonck, ‘An evaluation of the weighting procedures for an online access panel survey’, *Surv Res Methods*, 2008, doi: 10.18148/srm/2008.v2i2.82.

[145] J. A. Dever, A. Rafferty, and R. Valliant, ‘Internet surveys: Can statistical adjustments eliminate coverage bias?’, *Surv Res Methods*, 2008, doi: 10.18148/srm/2008.v2i2.128.

[146] S. Biffignandi and J. Bethlehem, ‘Sampling for Web Surveys’, in *Handbook of Web Surveys*, 2021. doi: 10.1002/9781119371717.ch4.

[147] S. Lee and R. Valliant, ‘Estimation for volunteer panel web surveys using propensity score adjustment and calibration adjustment’, *Sociol Methods Res*, 2009, doi: 10.1177/0049124108329643.

[148] S. V. Stehman, C. C. Fonte, G. M. Foody, and L. See, ‘Using volunteered geographic information (VGI) in design-based statistical inference for area estimation and accuracy assessment of land cover’, *Remote Sens Environ*, 2018, doi: 10.1016/j.rse.2018.04.014.

[149] D. M. Chen and H. Wei, ‘The effect of spatial autocorrelation and class proportion on the accuracy measures from different sampling designs’, *ISPRS Journal of Photogrammetry and Remote Sensing*, 2009, doi: 10.1016/j.isprsjprs.2008.07.004.

[150] J. B. Campbell, ‘Spatial correlation effects upon accuracy of supervised classification of land cover.’, *Photogramm Eng Remote Sensing*, 1981.

[151] J. S. Ay, R. Chakir, and J. Le Gallo, ‘Aggregated Versus Individual Land-Use Models: Modeling Spatial Autocorrelation to Increase Predictive Accuracy’, *Environmental Modeling and Assessment*, 2017, doi: 10.1007/s10666-016-9523-5.

[152] J. D. Opsomer, M. Francisco-Fernández, and X. Li, ‘Model-Based Non-parametric Variance Estimation for Systematic Sampling’, *Scandinavian Journal of Statistics*, 2012, doi: 10.1111/j.1467-9469.2011.00773.x.

[153] S. Magnussen and T. Nord-Larsen, ‘Design-consistent model-based variances with systematic sampling: a case study with the Danish national Forest inventory’, *Commun Stat Simul Comput*, 2021, doi: 10.1080/03610918.2018.1547401.

[154] K. M. Wolter, ‘Variance Estimation for Systematic Sampling’, in *Variance Estimation for Systematic Sampling*, Springer, New York, NY, 2008, pp. 289–353. doi: https://doi.org/10.1007/978-0-387-35099-8_8.

[155] G. H. Strand, ‘A study of variance estimation methods for systematic spatial sampling’, *Spat Stat*, 2017, doi: 10.1016/j.spasta.2017.06.008.

[156] Y. Alimohamadi and M. Sepandi, ‘Considering the design effect in cluster sampling’, *J Cardiovasc Thorac Res*, 2019, doi: 10.15171/jcvtr.2019.14.

[157] S. V. Stehman, ‘Basic probability sampling designs for thematic map accuracy assessment’, *Int J Remote Sens*, 1999, doi: 10.1080/014311699212100.

[158] K. M. Bretthauer, A. Ross, and B. Shetty, ‘Nonlinear integer programming for optimal allocation in stratified sampling’, *Eur J Oper Res*, 1999, doi: 10.1016/S0377-2217(98)00180-5.

[159] M. G. M. Khan, E. A. Khan, and M. J. Ahsan, ‘An optimal multivariate stratified sampling design using dynamic programming’, *Aust N Z J Stat*, 2003, doi: 10.1111/1467-842X.00264.

[160] M. G. M. Khan, T. Maiti, and M. J. Ahsan, ‘An optimal multivariate stratified sampling design using auxiliary information: An integer solution using goal programming approach’, *J Off Stat*, 2010.

[161] S. Gabler, M. Ganninger, and R. Münnich, ‘Optimal allocation of the sample size to strata under box constraints’, *Metrika*, 2012, doi: 10.1007/s00184-010-0319-3.

[162] S. Khowaja, S. Ghufran, and M. J. Ahsan, ‘Estimation of population means in multivariate stratified random sampling’, *Commun Stat Simul Comput*, 2011, doi: 10.1080/03610918.2010.551014.

[163] J. Buddhakulsomsiri and P. Parthanadee, ‘Stratified random sampling for estimating billing accuracy in health care systems’, *Health Care Manag Sci*, 2008, doi: 10.1007/s10729-007-9023-x.

[164] J. E. Wagner and S. V. Stehman, ‘Optimizing sample size allocation to strata for estimating area and map accuracy’, *Remote Sens Environ*, 2015, doi: 10.1016/j.rse.2015.06.027.

[165] P. Robinson and C. Sarndal, ‘Asymptotic properties of the generalized regression estimator in probability sampling’, *Sankhya : the Indian Journal of Statistics Series B*, 1983.

[166] C. E. Särndal, ‘On π-inverse weighting versus best linear unbiased weighting in probability sampling’, *Biometrika*, 1980, doi: 10.1093/biomet/67.3.639.

[167] Y. Tillé and M. Wilhelm, ‘Probability sampling designs: Principles for choice of design and balancing’, *Statistical Science*, 2017, doi: 10.1214/16-STS606.

[168] J. Kiefer, ‘Optimum Experimental Designs’, *Journal of the Royal Statistical Society: Series B (Methodological),* 1959, doi: 10.1111/j.2517-6161.1959.tb00338.x.

[169] A. C. Atkinson, ‘The Usefulness of Optimum Experimental Designs’, *Journal of the Royal Statistical Society: Series B (Methodological)*, 1996, doi: 10.1111/j.2517-6161.1996.tb02067.x.

[170] C. M. Anderson-Cook, ‘Optimum Experimental Designs, With SAS’, *J Am Stat Assoc,* 2008, doi: 10.1198/jasa.2008.s258.

[171] I. Bauer, H. G. Bock, S. Körkel, and J. P. Schlöder, ‘Numerical methods for optimum experimental design in DAE systems’, *J Comput Appl Math,* 2000, doi: 10.1016/S0377-0427(00)00300-9.

[172] J. R. Banga and E. Balsa-Canto, ‘Parameter estimation and optimal experimental design’, *Essays Biochem*, 2008, doi: 10.1042/BSE0450195.

[173] A. Alexanderian, ‘Optimal experimental design for infinite-dimensional Bayesian inverse problems governed by PDEs: A review’, *Inverse Problems*. 2021. doi: 10.1088/1361-6420/abe10c.

[174] E. Hofer, ‘When to separate uncertainties and when not to separate’, *Reliab Eng Syst Saf,* 1996, doi: 10.1016/S0951-8320(96)00068-3.

[175] E. Hüllermeier and W. Waegeman, ‘Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods’, *Mach Learn*, 2021, doi: 10.1007/s10994-021-05946-3.

[176] R. Senge *et al.*, ‘Reliable classification: Learning classifiers that distinguish aleatoric and epistemic uncertainty’, *Inf Sci (N Y),* 2014, doi: 10.1016/j.ins.2013.07.030.

[177] M. H. Shaker and E. Hüllermeier, ‘Aleatoric and Epistemic Uncertainty with Random Forests’, in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2020. doi: 10.1007/978-3-030-44584-3_35.

[178] A. Kendall and Y. Gal, ‘What uncertainties do we need in Bayesian deep learning for computer vision?’, in *Advances in Neural Information Processing Systems*, 2017.

[179] G. Li, L. Yang, C. G. Lee, X. Wang, and M. Rong, ‘A Bayesian Deep Learning RUL Framework Integrating Epistemic and Aleatoric Uncertainties’, *IEEE Transactions on Industrial Electronics*, 2021, doi: 10.1109/TIE.2020.3009593.

[180] E. R. Ziegel, E. L. Lehmann, and G. Casella, ‘Theory of Point Estimation’, *Technometrics*, 1999, doi: 10.2307/1270597.

[181] J. Cohen, ‘Statistical power analysis for the behavioural sciences. Hillside’, *NJ: Lawrence Earlbaum Associates*. 1988.

[182] M. Esteban-Bravo, A. Leszkiewicz, and J. M. Vidal-Sanz, ‘Exact optimal experimental designs with constraints’, *Stat Comput*, 2017, doi: 10.1007/s11222-016-9658-x.

[183] J. Vanlier, C. A. Tiemann, P. A. J. Hilbers, and N. A. W. van Riel, ‘A Bayesian approach to targeted experiment design’, *Bioinformatics*, 2012, doi: 10.1093/bioinformatics/bts092.

[184] E. F. Murphy, S. G. Gilmour, and M. J. C. Crabbe, ‘Efficient and accurate experimental design for enzyme kinetics: Bayesian studies reveal a systematic approach’, *J Biochem Biophys Methods*, 2003, doi: 10.1016/S0165-022X(02)00183-5.

[185] X. Huan and Y. M. Marzouk, ‘Gradient-based stochastic optimization methods in Bayesian experimental design’, *Int J Uncertain Quantif*, 2014, doi: 10.1615/Int.J.UncertaintyQuantification.2014006730.

[186] A. G. Carlon, B. M. Dia, L. Espath, R. H. Lopez, and R. Tempone, ‘Nesterov-aided stochastic gradient methods using Laplace approximation for Bayesian design optimization’, *Comput Methods Appl Mech Eng*, 2020, doi: 10.1016/j.cma.2020.112909.

[187] D. Fouskakis and D. Draper, ‘Stochastic optimization: A review’, *International Statistical Review*. 2002. doi: 10.1111/j.1751-5823.2002.tb00174.x.

[188] D. P. Kingma and J. L. Ba, ‘Adam: A method for stochastic optimization’, in *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 2015.

[189] A. Zakaria, F. B. Ismail, M. S. H. Lipu, and M. A. Hannan, ‘Uncertainty models for stochastic optimization in renewable energy applications’, *Renewable Energy*. 2020. doi: 10.1016/j.renene.2019.07.081.

[190] S. Russel and P. Norvig, *Artificial intelligence—a modern approach 3rd Edition*. 2012. doi: 10.1017/S0269888900007724.

[191] L. Rokach, ‘Ensemble-based classifiers’, *Artif Intell Rev*, 2010, doi: 10.1007/s10462-009-9124-7.

[192] O. Sagi and L. Rokach, ‘Ensemble learning: A survey’, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*. 2018. doi: 10.1002/widm.1249.

[193] N. Friedman, D. Geiger, and M. Goldszmidt, ‘Bayesian Network Classifiers’, *Mach Learn*, 1997, doi: 10.1023/a:1007465528199.

[194] I. Rish, ‘An empirical study of the naive Bayes classifier’, *IJCAI 2001 Work. Empir. methods Artif. Intell.*, 2001, doi: 10.1039/b104835j.

[195] Z. Zhang, ‘Introduction to machine learning: K-nearest neighbors’, *Ann Transl Med*, 2016, doi: 10.21037/atm.2016.03.37.

[196] M. Pesaresi, V. Syrris, and A. Julea, ‘A new method for earth observation data analytics based on symbolic machine learning’, *Remote Sens (Basel)*, 2016, doi: 10.3390/rs8050399.

[197] F. Neri and L. Saitta, ‘Exploring the power of genetic search in learning symbolic classifiers’, *IEEE Trans Pattern Anal Mach Intell*, 1996, doi: 10.1109/34.544085.

[198] A. J. Scott, D. W. Hosmer, and S. Lemeshow, ‘Applied Logistic Regression.’, *Biometrics*, 1991, doi: 10.2307/2532419.

[199] R. E. Neapolitan and X. Jiang, ‘Neural Networks and Deep Learning’, in *Artificial Intelligence*, 2018. doi: 10.1201/b22400-15.

[200] J. Schmidhuber, ‘Deep Learning in neural networks: An overview’, *Neural Networks*. 2015. doi: 10.1016/j.neunet.2014.09.003.

[201] R. G. Brereton and G. R. Lloyd, ‘Support Vector Machines for classification and regression’, *Analyst*. 2010. doi: 10.1039/b918972f.

[202] M. Seeger, ‘Gaussian processes for machine learning.’, *International journal of neural systems*. 2004. doi: 10.1142/S0129065704001899.

[203] S. N. Wood, *Generalized additive models: An introduction with R, second edition*. 2017. doi: 10.1201/9781315370279.

[204] J. H. Friedman and C. B. Roosen, ‘An introduction to multivariate adaptive regression splines’, *Stat Methods Med Res*, 1995, doi: 10.1177/096228029500400303.

[205] K. R. Müller, S. Mika, G. Rätsch, K. Tsuda, and B. Schölkopf, ‘An introduction to kernel-based learning algorithms’, *IEEE Transactions on Neural Networks*. 2001. doi: 10.1109/72.914517.

[206] R. E. Schapire, ‘Explaining adaboost’, in *Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik*, 2013. doi: 10.1007/978-3-642-41136-6_5.

[207] A. Natekin and A. Knoll, ‘Gradient boosting machines, a tutorial’, *Front Neurorobot*, 2013, doi: 10.3389/fnbot.2013.00021.

[208] Z. P. Brodeur, J. D. Herman, and S. Steinschneider, ‘Bootstrap Aggregation and Cross-Validation Methods to Reduce Overfitting in Reservoir Control Policy Search’, *Water Resour Res*, 2020, doi: 10.1029/2020WR027184.

[209] S. Sun and C. Zhang, ‘Subspace ensembles for classification’, *Physica A: Statistical Mechanics and its Applications*, vol. 385, no. 1, pp. 199–207, Nov. 2007, doi: 10.1016/J.PHYSA.2007.05.010.

[210] T. G. Dietterich and G. Bakiri, ‘Solving Multiclass Learning Problems via Error-Correcting Output Codes’, *Journal of Artificial Intelligence Research*, vol. 2, pp. 263–286, Jan. 1994, doi: 10.1613/JAIR.105.

[211] R. Bryll, R. Gutierrez-Osuna, and F. Quek, ‘Attribute bagging: improving accuracy of classifier ensembles by using random feature subsets’, *Pattern Recognit*, vol. 36, no. 6, pp. 1291–1302, Jun. 2003, doi: 10.1016/S0031-3203(02)00121-8.

[212] L. Breiman, ‘Random forests’, *Mach Learn*, 2001, doi: 10.1023/A:1010933404324.

[213] J. E. van Engelen and H. H. Hoos, ‘A survey on semi-supervised learning’, *Mach Learn*, 2020, doi: 10.1007/s10994-019-05855-6.

[214] O. L. Thabeng, S. Merlo, and E. Adam, ‘High-resolution remote sensing and advanced classification techniques for the prospection of archaeological sites’ markers: The case of dung deposits in the Shashi-Limpopo Confluence area (southern Africa)’, *J Archaeol Sci*, 2019, doi: 10.1016/j.jas.2018.12.003.

[215] A. Rasmus, H. Valpola, M. Honkala, M. Berglund, and T. Raiko, ‘Semi-supervised learning with Ladder networks’, in *Advances in Neural Information Processing Systems*, 2015.

[216] C. Leistner, A. Saffari, J. Santner, and H. Bischof, ‘Semi-supervised random forests’, in *Proceedings of the IEEE International Conference on Computer Vision*, 2009. doi: 10.1109/ICCV.2009.5459198.

[217] N. D. Lawrence and M. I. Jordan, ‘Semi-supervised Learning via gaussian processes’, in *Advances in Neural Information Processing Systems*, 2005.

[218] B. Frénay and M. Verleysen, ‘Classification in the presence of label noise: A survey’, *IEEE Trans Neural Netw Learn Syst*, 2014, doi: 10.1109/TNNLS.2013.2292894.

[219] T. Xiao, T. Xia, Y. Yang, C. Huang, and X. Wang, ‘Learning from massive noisy labeled data for image classification’, *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 07-12-June-2015, pp. 2691–2699, Oct. 2015, doi: 10.1109/CVPR.2015.7298885.

[220] D. Song, C. Lee, Y. Li, and D. S. Neural, ‘Blind regression: Nonparametric regression for latent variable models via collaborative filtering’, *Adv Neural Inf Process Syst*, vol. 29, 2016, Accessed: Jan. 02, 2023. [Online]. Available: https://proceedings.neurips.cc/paper/2016/hash/678a1491514b7f1006d605e9161946b1-Abstract.html

[221] F. Murtagh and P. Contreras, ‘Algorithms for hierarchical clustering: an overview, II’, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*. 2017. doi: 10.1002/widm.1219.

[222] M. Ahmed, R. Seraj, and S. M. S. Islam, ‘The k-means algorithm: A comprehensive survey and performance evaluation’, *Electronics (Switzerland)*. 2020. doi: 10.3390/electronics9081295.

[223] M. Hahsler, M. Piekenbrock, and D. Doran, ‘Dbscan: Fast density-based clustering with R’, *J Stat Softw*, 2019, doi: 10.18637/jss.v091.i01.

[224] H. Gan, N. Sang, R. Huang, X. Tong, and Z. Dan, ‘Using clustering analysis to improve semi-supervised classification’, *Neurocomputing*, vol. 101, pp. 290–298, Feb. 2013, doi: 10.1016/J.NEUCOM.2012.08.020.

[225] N. M. N. Mathivanan, N. A. Nor, and R. M. Janor, ‘Improving Classification Accuracy Using Clustering Technique’, *Bulletin of Electrical Engineering and Informatics*, vol. 7, no. 3, pp. 465–470, Sep. 2018, doi: 10.11591/EEI.V7I3.1272.

[226] R. Dara, S. C. Kremer, and D. A. Stacey, ‘Clustering unlabeled data with SOMs improves classification of labeled real-world data’, *Proceedings of the International Joint Conference on Neural Networks*, vol. 3, pp. 2237–2242, 2002, doi: 10.1109/IJCNN.2002.1007489.

[227] M. Peikari, S. Salama, S. Nofech-Mozes, and A. L. Martel, ‘A Cluster-then-label Semi-supervised Learning Approach for Pathology Image Classification’, *Scientific Reports 2018 8:1*, vol. 8, no. 1, pp. 1–13, May 2018, doi: 10.1038/s41598-018-24876-0.

[228] M. Köppen, ‘The curse of dimensionality’, *5th Online World Conf. Soft Comput. Ind. Appl.*, vol. 1, pp. 4–8, 2000.

[229] H. Abdi and L. J. Williams, ‘Principal component analysis’, *Wiley Interdiscip Rev Comput Stat*, vol. 2, no. 4, pp. 433–459, Jul. 2010, doi: 10.1002/WICS.101.

[230] T. Kohonen, ‘The self-organizing map’, *Neurocomputing*, 1998, doi: 10.1016/S0925-2312(98)00030-7.

[231] W. Wang, Y. Huang, Y. Wang, and L. Wang, ‘Generalized autoencoder: A neural network framework for dimensionality reduction’, *IEEE Computer Society Conference*

on Computer Vision and Pattern Recognition Workshops, pp. 496–503, Sep. 2014, doi: 10.1109/CVPRW.2014.79.

[232] V. S. Sheng, F. Provost, and P. G. Ipeirotis, ‘Get another label? Improving data quality and data mining using multiple, noisy labelers’, in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2008. doi: 10.1145/1401890.1401965.

[233] E. Barsoum, C. Zhang, C. C. Ferrer, and Z. Zhang, ‘Training deep networks for facial expression recognition with crowd-sourced label distribution’, *ICMI 2016 - Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pp. 279–283, Oct. 2016, doi: 10.1145/2993148.2993165.

[234] V. C. Raykar *et al.*, ‘Supervised learning from multiple experts : Whom to trust when everyone lies a bit’, *ACM International Conference Proceeding Series*, vol. 382, 2009, doi: 10.1145/1553374.1553488.

[235] F. Muhlenbach, S. Lallich, and D. A. Zighed, ‘Identifying and Handling Mislabelled Instances’, in *Journal of Intelligent Information Systems*, 2004. doi: 10.1023/A:1025832930864.

[236] Z. Zhou, ‘Multi-instance learning: a survey’, *AI Lab, Dep. Comput. Sci. Technol.*, 2004.

[237] J. Foulds and E. Frank, ‘A review of multi-instance learning assumptions’, *Knowledge Engineering Review*. 2010. doi: 10.1017/S026988890999035X.

[238] Z. H. Zhou, ‘A brief introduction to weakly supervised learning’, *Natl. Sci. Rev.* 2018. doi: 10.1093/nsr/nwx106.

[239] M. A. Carbonneau, V. Cheplygina, E. Granger, and G. Gagnon, ‘Multiple instance learning: A survey of problem characteristics and applications’, *Pattern Recognit.*, 2018, doi: 10.1016/j.patcog.2017.10.009.

[240] F. Zhuang *et al.*, ‘A Comprehensive Survey on Transfer Learning’, *Proceedings of the IEEE*. 2021. doi: 10.1109/JPROC.2020.3004555.

[241] K. Weiss, T. M. Khoshgoftaar, and D. D. Wang, ‘A survey of transfer learning’, *J Big Data*, 2016, doi: 10.1186/s40537-016-0043-6.

[242] S. J. Pan and Q. Yang, ‘A survey on transfer learning’, *IEEE Transactions on Knowledge and Data Engineering*. 2010. doi: 10.1109/TKDE.2009.191.

[243] H. I. Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P. A. Muller, ‘Transfer learning for time series classification’, *Proceedings - 2018 IEEE International Conference on Big Data, Big Data 2018*, pp. 1367–1376, Jan. 2019, doi: 10.1109/BIGDATA.2018.8621990.

[244] S. Ruder, M. E. Peters, S. Swayamdipta, and T. Wolf, ‘Transfer Learning in Natural Language Processing’, *Proceedings of the 2019 Conference of the North*, pp. 15–18, 2019, doi: 10.18653/V1/N19-5004.

[245] C. Raffel *et al.*, ‘Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer’, *Journal of Machine Learning Research*, vol. 21, pp. 1–67, 2020, Accessed: Jan. 07, 2023. [Online]. Available: http://jmlr.org/papers/v21/20-074.html.

[246] M. Shaha and M. Pawar, ‘Transfer Learning for Image Classification’, *Proceedings of the 2nd International Conference on Electronics, Communication and Aerospace Technology, ICECA 2018*, pp. 656–660, Sep. 2018, doi: 10.1109/ICECA.2018.8474802.

[247] M. Hussain, J. J. Bird, and D. R. Faria, ‘A study on CNN transfer learning for image classification’, *Advances in Intelligent Systems and Computing*, vol. 840, pp. 191–202, 2019, doi: 10.1007/978-3-319-97982-3_16/COVER.

[248] F. Provost, ‘Machine learning from imbalanced data sets 101’, *Proceedings of the AAAI'2000 Workshop on …*, 2000.

[249] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, ‘SMOTE: Synthetic minority over-sampling technique’, *Journal of Artificial Intelligence Research*, 2002.

[250] Y. Qian *et al.*, ‘Adversarial Example Generation Based on Particle Swarm Optimization’, *Dianzi Yu Xinxi Xuebao/Journal of Electronics and Information Technology*, 2019, doi: 10.11999/JEITdzyxxxb-41-7-1658.

[251] D. Wang, L. Dong, R. Wang, D. Yan, and J. Wang, ‘Targeted Speech Adversarial Example Generation with Generative Adversarial Network’, *IEEE Access*, 2020, doi: 10.1109/ACCESS.2020.3006130.

[252] D. Malmgren-Hansen, A. Kusk, J. Dall, A. A. Nielsen, R. Engholm, and H. Skriver, ‘Improving SAR Automatic Target Recognition Models with Transfer Learning from Simulated Data’, *IEEE Geoscience and Remote Sensing Letters*, 2017, doi: 10.1109/LGRS.2017.2717486.

[253] H. Tercan, A. Guajardo, J. Heinisch, T. Thiele, C. Hopmann, and T. Meisen, ‘Transfer-Learning: Bridging the Gap between Real and Simulation Data for Machine Learning in Injection Molding’, in *Procedia CIRP*, 2018. doi: 10.1016/j.procir.2018.03.087.

[254] D. Bobylev, T. Choudhury, J. O. Miettinen, R. Viitala, E. Kurvinen, and J. Sopanen, ‘Simulation-Based Transfer Learning for Support Stiffness Identification’, *IEEE Access*, 2021, doi: 10.1109/ACCESS.2021.3108414.

[255] S. Zhang, G. Yang, T. Sun, K. Du, and J. Guo, ‘Uav detection with transfer learning from simulated data of laser active imaging’, *Applied Sciences (Switzerland)*, 2021, doi: 10.3390/app11115182.

[256] C. Schaffer, ‘Selecting a classification method by cross-validation’, *Mach Learn*, vol. 13, no. 1, pp. 135–143, Oct. 1993, doi: 10.1007/BF00993106.

[257] B. Ramosaj and M. Pauly, ‘Consistent estimation of residual variance with random forest Out-Of-Bag errors’, *Stat Probab Lett*, vol. 151, pp. 49–57, Aug. 2019, doi: 10.1016/J.SPL.2019.03.017.

[258] D. Berrar and J. A. Lozano, ‘Significance tests or confidence intervals: which are preferable for the comparison of classifiers?’, *http://dx.doi.org/10.1080/0952813X.2012.680252*, vol. 25, no. 2, pp. 189–206, Jun. 2013, doi: 10.1080/0952813X.2012.680252.

[259] J. Demšar, ‘Statistical Comparisons of Classifiers over Multiple Data Sets’, *Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.

[260] G. Corani, A. Benavoli, J. Demšar, F. Mangili, and M. Zaffalon, ‘Statistical comparison of classifiers through Bayesian hierarchical modelling’, *Mach Learn*, vol. 106, no. 11, pp. 1817–1837, Nov. 2017, doi: 10.1007/S10994-017-5641-9/FIGURES/7.

[261] M. I. Faisal, S. Bashir, Z. S. Khan, and F. Hassan Khan, ‘An Evaluation of Machine Learning Classifiers and Ensembles for Early Stage Prediction of Lung Cancer’, in *2018 3rd International Conference on Emerging Trends in Engineering, Sciences and Technology, ICEEST 2018*, 2019. doi: 10.1109/ICEEST.2018.8643311.

[262] J. Singh and J. Singh, ‘A survey on machine learning-based malware detection in executable files’, *Journal of Systems Architecture*. 2021. doi: 10.1016/j.sysarc.2020.101861.

[263] J. Vanschoren, ‘Meta-learning’, in *Automated Machine Learning*, 2019, pp. 35–61.

[264] T. Hospedales, A. Antoniou, P. Micaelli, and A. Storkey, ‘Meta-Learning in Neural Networks: A Survey’, *IEEE Trans Pattern Anal Mach Intell*, vol. 44, no. 9, pp. 5149–5169, Sep. 2022, doi: 10.1109/TPAMI.2021.3079209.

[265] M. Huisman, J. N. van Rijn, and A. Plaat, ‘A survey of deep meta-learning’, *Artif Intell Rev*, vol. 54, no. 6, pp. 4483–4541, Aug. 2021, doi: 10.1007/S10462-021-10004-4/TABLES/5.

[266] P. Sibi, S. Allwyn Jones, and P. Siddarth, ‘Analysis of different activation functions using back propagation neural networks’, *J Theor Appl Inf Technol*, 2013.

[267] S. Sharma, S. Sharma, and A. Anidhya, ‘Understanding Activation Functions in Neural Networks’, *International Journal of Engineering Applied Sciences and Technology*, 2020.

[268] R. G. Negri, E. A. Da Silva, and W. Casaca, ‘Inducing Contextual Classifications with Kernel Functions into Support Vector Machines’, *IEEE Geoscience and Remote Sensing Letters*, 2018, doi: 10.1109/LGRS.2018.2816460.

[269] S. Amari and S. Wu, ‘Improving support vector machine classifiers by modifying kernel functions’, *Neural Networks*, 1999, doi: 10.1016/S0893-6080(99)00032-5.

[270] N. V. Chawla and G. Karakoulas, ‘Learning from labeled and unlabeled data: An empirical study across techniques and domains’, *Journal of Artificial Intelligence Research*, 2005, doi: 10.1613/jair.1509.

[271] Y. F. Li and Z. H. Zhou, ‘Towards making unlabeled data never hurt’, *IEEE Trans Pattern Anal Mach Intell*, 2015, doi: 10.1109/TPAMI.2014.2299812.

[272] Y. F. Li, J. T. Kwok, and Z. H. Zhou, ‘Towards safe semi-supervised learning for multivariate performance measures’, in *30th AAAI Conference on Artificial Intelligence, AAAI 2016*, 2016.

[273] A. Gaba and R. L. Winkler, ‘Implications of Errors in Survey Data: A Bayesian Model’, *Manage Sci*, 1992, doi: 10.1287/mnsc.38.7.913.

[274] R. J. Hickey, ‘Noise modelling and evaluating learning from examples’, *Artif Intell*, 1996, doi: 10.1016/0004-3702(94)00094-8.

[275] M. A. Carbonneau, V. Cheplygina, E. Granger, and G. Gagnon, ‘Multiple instance learning: A survey of problem characteristics and applications’, *Pattern Recognit*, 2018, doi: 10.1016/j.patcog.2017.10.009.

[276] A. Argyriou, A. Maurer, and M. Pontil, ‘An algorithm for transfer learning in a heterogeneous environment’, in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2008. doi: 10.1007/978-3-540-87479-9_23.

[277] L. Ge, J. Gao, H. Ngo, K. Li, and A. Zhang, ‘On handling negative transfer and imbalanced distributions in multiple source transfer learning’, *Stat Anal Data Min*, 2014, doi: 10.1002/sam.11217.

[278] B. Bakker and T. Heskes, ‘Task clustering and gating for bayesian multitask learning’, *Journal of Machine Learning Research*, 2004, doi: 10.1162/153244304322765658.

[279] K. A. Smith-Miles, ‘Cross-disciplinary perspectives on meta-learning for algorithm selection’, *ACM Computing Surveys (CSUR)*, vol. 41, no. 1, Jan. 2009, doi: 10.1145/1456650.1456656.

[280] M. Maher and S. Sakr, ‘SmartML: A Meta Learning-Based Framework for Automated Selection and Hyperparameter Tuning for Machine Learning Algorithms’, *Advances in Database Technology - EDBT*, vol. 2019-March, pp. 554–557, Mar. 2019, doi: 10.5441/002/EDBT.2019.54.

[281] J. P. Monteiro, D. Ramos, D. Carneiro, F. Duarte, J. M. Fernandes, and P. Novais, ‘Meta-learning and the new challenges of machine learning’, *International Journal of Intelligent Systems*, vol. 36, no. 11, pp. 6240–6272, Nov. 2021, doi: 10.1002/INT.22549.

[282] G. E. A. P. A. Batista and M. C. Monard, ‘An analysis of four missing data treatment methods for supervised learning’, *Applied Artificial Intelligence*, 2003, doi: 10.1080/713827181.

[283] B. M. Marlin, ‘Missing Data Problems in Machine Learning’, *Graduate Department of Computer Science*, 2008.

[284] H. Osman, M. Ghafari, and O. Nierstrasz, ‘Hyperparameter optimization to improve bug prediction accuracy’, in *MaLTeSQuE 2017 - IEEE International Workshop on Machine Learning Techniques for Software Quality Evaluation, co-located with SANER 2017*, 2017. doi: 10.1109/MALTESQUE.2017.7882014.

[285] P. Schratz, J. Muenchow, E. Iturritxa, J. Richter, and A. Brenning, ‘Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data’, *Ecol Modell*, 2019, doi: 10.1016/j.ecolmodel.2019.06.002.

[286] M. J. Denny and A. Spirling, ‘Text Preprocessing for Unsupervised Learning: Why It Matters, When It Misleads, and What to Do about It’, *Political Analysis*, 2018, doi: 10.1017/pan.2017.44.

[287] E. Law *et al.*, ‘The Science of Citizen Science’, 2017. doi: 10.1145/3022198.3022652.

[288] K. Lambers, W. B. Verschoof-van der Vaart, and Q. P. J. Bourgeois, ‘Integrating remote sensing, machine learning, and citizen science in dutch archaeological prospection’, *Remote Sens (Basel)*, 2019, doi: 10.3390/rs11070794.

[289] K. Sparks, A. Klippel, J. O. Wallgrün, and D. Mark, ‘Citizen science land cover classification based on ground and aerial imagery’, in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2015. doi: 10.1007/978-3-319-23374-1_14.

[290] K. A. Lee, J. R. Lee, and P. Bell, ‘A review of Citizen Science within the Earth Sciences: potential benefits and obstacles’, *Proceedings of the Geologists' Association*. 2020. doi: 10.1016/j.pgeola.2020.07.010.

[291] J. C. L. Bayas *et al.*, ‘Crowdsourcing in-situ data on land cover and land use using gamification and mobile technology’, *Remote Sens (Basel)*, 2016, doi: 10.3390/rs8110905.

[292] B. Settles, ‘Active Learning Literature Survey’, *Mach Learn*, 2010, doi: 10.1.1.167.4245.

[293] D. A. Cohn, Z. Ghahramani, and M. I. Jordan, ‘Active learning with statistical models’, *Journal of Artificial Intelligence Research*, 1996, doi: 10.1613/jair.295.

[294] R. M. Castro and R. D. Nowak, ‘Minimax bounds for active learning’, *IEEE Trans Inf Theory*, 2008, doi: 10.1109/TIT.2008.920189.

[295] P. Ren *et al.*, ‘A Survey of Deep Active Learning’, *ACM Comput Surv*, vol. 54, no. 9, Dec. 2022, doi: 10.1145/3472291.

[296] G. Cavallanti, N. Cesa-Bianchi, and C. Gentile, ‘Learning noisy linear classifiers via adaptive and selective sampling’, *Mach Learn*, 2011, doi: 10.1007/s10994-010-5191-x.

[297] A. Basudhar and S. Missoum, ‘An improved adaptive sampling scheme for the construction of explicit boundaries’, *Structural and Multidisciplinary Optimization*, 2010, doi: 10.1007/s00158-010-0511-0.

[298] D. Needell, N. Srebro, and R. Ward, ‘Stochastic gradient descent, weighted sampling, and the randomized Kaczmarz algorithm’, *Math Program*, 2016, doi: 10.1007/s10107-015-0864-7.

[299] X. Peng, L. Li, and F. Y. Wang, ‘Accelerating Minibatch Stochastic Gradient Descent Using Typicality Sampling’, *IEEE Trans Neural Netw Learn Syst*, 2020, doi: 10.1109/TNNLS.2019.2957003.

[300] J. I. Avalos-López, A. Rojas-Domínguez, M. Ornelas-Rodríguez, M. Carpio, and S. I. Valdez, ‘Efficient Training of Deep Learning Models Through Improved Adaptive Sampling’, in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2021. doi: 10.1007/978-3-030-77004-4_14.

[301] Y. Gal, R. Islam, and Z. Ghahramani, ‘Deep Bayesian active learning with image data’, in *34th International Conference on Machine Learning, ICML 2017*, 2017. doi: 10.17863/CAM.11070.

[302] A. Kirsch, J. van Amersfoort, and Y. Gal, ‘BatchBALD: Efficient and diverse batch acquisition for deep Bayesian active learning’, in *Advances in Neural Information Processing Systems*, 2019.

[303] T. Lookman, P. V. Balachandran, D. Xue, and R. Yuan, ‘Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design’, *npj Computational Materials*. 2019. doi: 10.1038/s41524-019-0153-8.

[304] B. Miller, F. Linder, and W. R. Mebane, ‘Active learning approaches for labeling text: Review and assessment of the performance of active learning approaches’, *Political Analysis*. 2020. doi: 10.1017/pan.2020.4.

[305] Y. Shen, H. Yun, Z. C. Lipton, Y. Kronrod, and A. Anandkumar, ‘Deep Active Learning for Named Entity Recognition’, *Proceedings of the 2nd Workshop on Representation Learning for NLP, Rep4NLP 2017 at the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017*, pp. 252–256, Jul. 2017, doi: 10.48550/arxiv.1707.05928.

[306] Y. Li, B. Fan, W. Zhang, W. Ding, and J. Yin, ‘Deep active learning for object detection’, *Inf Sci (N Y)*, vol. 579, pp. 418–433, Nov. 2021, doi: 10.1016/J.INS.2021.08.019.

[307] F. Scheibein, W. Donnelly, and J. S. Wells, ‘Assessing open science and citizen science in addictions and substance use research: A scoping review’, *International Journal of Drug Policy*. 2022. doi: 10.1016/j.drugpo.2021.103505.

[308] T. van Gestel, J. A. K. Suykens, G. Lanckriet, A. Lambrechts, B. de Moor, and J. Vandewalle, ‘Bayesian Framework for Least-Squares Support Vector Machine Classifiers, Gaussian Processes, and Kernel Fisher Discriminant Analysis’, *Neural Comput*, vol. 14, no. 5, pp. 1115–1147, May 2002, doi: 10.1162/089976602753633411.

[309] P. Sollich, ‘Bayesian methods for support vector machines: Evidence and predictive class probabilities’, *Mach Learn*, vol. 46, no. 1–3, pp. 21–52, Jan. 2002, doi: 10.1023/A:1012489924661/METRICS.

[310] W. Hao and D. Y. Yeung, ‘Towards Bayesian Deep Learning: A Framework and Some Existing Methods’, *IEEE Trans Knowl Data Eng*, 2016, doi: 10.1109/TKDE.2016.2606428.

[311] F. Wenzel, T. Galy-Fajou, M. Deutsch, and M. Kloft, ‘Bayesian Nonlinear Support Vector Machines for Big Data’, *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10534 LNAI, pp. 307–322, 2017, doi: 10.1007/978-3-319-71249-9_19/FIGURES/3.

[312] J. Hensman, N. Fusi, and N. D. Lawrence, ‘Gaussian Processes for Big Data’, *Uncertainty in Artificial Intelligence - Proceedings of the 29th Conference, UAI 2013*, pp. 282–290, Sep. 2013, doi: 10.48550/arxiv.1309.6835.

[313] J. Hensman, A. G. Matthews, and Z. Ghahramani, ‘Scalable Variational Gaussian Process Classification’, *Journal of Machine Learning Research*, vol. 38, pp. 351–360, Nov. 2014, doi: 10.48550/arxiv.1411.2005.

[314] M. A. Newton, N. G. Polson, and J. Xu, ‘Weighted Bayesian bootstrap for scalable posterior distributions’, *Canadian Journal of Statistics*, 2021, doi: 10.1002/cjs.11570.

[315] K. Azizzadenesheli, E. Brunskill, and A. Anandkumar, ‘Efficient exploration through Bayesian deep Q-networks’, in *2018 Information Theory and Applications Workshop, ITA 2018*, 2018. doi: 10.1109/ITA.2018.8503252.

[316] M. E. Khan, D. Nielsen, V. Tangkaratt, W. Lin, Y. Gal, and A. Srivastava, ‘Fast and scalable Bayesian deep learning by weight-perturbation in Adam’, in *35th International Conference on Machine Learning, ICML 2018*, 2018.

[317] W. J. Maddox, T. Garipov, Izmailov, D. Vetrov, and A. G. Wilson, ‘A simple baseline for Bayesian uncertainty in deep learning’, in *Advances in Neural Information Processing Systems*, 2019.

[318] F. Doshi-Velez and B. Kim, ‘Towards A Rigorous Science of Interpretable Machine Learning’, Feb. 2017, Accessed: Jan. 09, 2023. [Online]. Available: http://arxiv.org/abs/1702.08608

[319] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, ‘Explainable ai: A review of machine learning interpretability methods’, *Entropy*, 2020, doi: 10.3390/e23010018.

[320] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, ‘Explaining explanations: An overview of interpretability of machine learning’, *Proceedings - 2018 IEEE 5th International Conference on Data Science and Advanced Analytics, DSAA 2018*, pp. 80–89, Jan. 2019, doi: 10.1109/DSAA.2018.00018.

[321] E. Borgonovo and E. Plischke, ‘Sensitivity analysis: A review of recent advances’, *Eur J Oper Res*, vol. 248, no. 3, pp. 869–887, Feb. 2016, doi: 10.1016/J.EJOR.2015.06.032.

[322] I. M. Sobol’ and S. Kucherenko, ‘Derivative based global sensitivity measures and their link with global sensitivity indices’, *Math Comput Simul*, vol. 79, no. 10, pp. 3009–3017, Jun. 2009, doi: 10.1016/J.MATCOM.2009.01.023.

[323] S. Janitza, E. Celik, and A. L. Boulesteix, ‘A computationally fast variable importance test for random forests for high-dimensional data’, *Adv Data Anal Classif*, vol. 12, no. 4, pp. 885–915, Jan. 2018, doi: 10.1007/S11634-016-0276-4/TABLES/3.

[324] T. A. Mara, ‘Extension of the RBD-FAST method to the computation of global sensitivity indices’, *Reliab Eng Syst Saf*, vol. 94, no. 8, pp. 1274–1281, Aug. 2009, doi: 10.1016/J.RESS.2009.01.012.

[325] J. Ish-Horowicz, K. Scharfstein, S. Flaxman, S. Filippi, D. Udwin, and L. Crawford, ‘Interpreting Deep Neural Networks Through Variable Importance’, *Journal of Machine Learning Research*, vol. 21, pp. 1–30, Jan. 2019, doi: 10.48550/arxiv.1901.09839.

[326] F. Campolongo, J. Cariboni, A. S.-E. modelling & software, and undefined 2007, ‘An effective screening design for sensitivity analysis of large models’, *Elsevier*, 2007, doi: 10.1016/j.envsoft.2006.10.004.

[327] I. J. Goodfellow, J. Shlens, and C. Szegedy, ‘Explaining and harnessing adversarial examples’, *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 2015.

[328] J. Su, D. V. Vargas, and K. Sakurai, ‘One Pixel Attack for Fooling Deep Neural Networks’, *IEEE Transactions on Evolutionary Computation*, vol. 23, no. 5, pp. 828–841, Oct. 2019, doi: 10.1109/TEVC.2019.2890858.

[329] J. Chen, X. Wu, Y. Guo, Y. Liang, and S. Jha, ‘Towards Evaluating the Robustness of Neural Networks Learned by Transduction’, Oct. 2021, Accessed: Jan. 22, 2023. [Online]. Available: http://arxiv.org/abs/2110.14735

[330] J. R. Zilke, E. L. Mencía, and F. Janssen, ‘DeepRED – Rule extraction from deep neural networks’, *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9956 LNAI, pp. 457–473, 2016, doi: 10.1007/978-3-319-46307-0_29/FIGURES/8.

[331] G. P. J. Schmitz, C. Aldrich, and F. S. Gouws, ‘ANN-DT: An algorithm for extraction of decision trees from artificial neural networks’, *IEEE Trans Neural Netw*, vol. 10, no. 6, pp. 1392–1401, 1999, doi: 10.1109/72.809084.

[332] M. T. Ribeiro, S. Singh, and C. Guestrin, '"Why should i trust you?" Explaining the predictions of any classifier', *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, vol. 13-17-August-2016, pp. 1135–1144, Aug. 2016, doi: 10.1145/2939672.2939778.

[333] S. M. Lundberg, P. G. Allen, and S.-I. Lee, 'A Unified Approach to Interpreting Model Predictions', *Adv Neural Inf Process Syst*, vol. 30, 2017, Accessed: Jan. 14, 2023. [Online]. Available: https://github.com/slundberg/shap

[334] M. Staniak and P. Biecek, 'Explanations of model predictions with live and breakDown packages', *R Journal*, vol. 10, no. 2, pp. 395–409, Apr. 2018, doi: 10.32614/RJ-2018-072.

[335] S. M. Mathews, 'Explainable Artificial Intelligence Applications in NLP, Biomedical, and Malware Classification: A Literature Review', *Advances in Intelligent Systems and Computing*, vol. 998, pp. 1269–1292, 2019, doi: 10.1007/978-3-030-22868-2_90/COVER.

[336] M. Danilevsky, K. Qian, R. Aharonov, Y. Katsis, B. Kawas, and P. Sen, 'A Survey of the State of Explainable AI for Natural Language Processing', Oct. 2020, doi: 10.48550/arxiv.2010.00711.

[337] C. Shi, L. Fang, Z. Lv, and M. Zhao, 'Explainable scale distillation for hyperspectral image classification', *Pattern Recognit*, vol. 122, p. 108316, Feb. 2022, doi: 10.1016/J.PATCOG.2021.108316.

[338] H. Lee *et al.*, 'An explainable deep-learning algorithm for the detection of acute intracranial haemorrhage from small datasets', *Nature Biomedical Engineering 2018 3:3*, vol. 3, no. 3, pp. 173–182, Dec. 2018, doi: 10.1038/s41551-018-0324-9.

[339] E. Pintelas, M. Liaskos, I. E. Livieris, S. Kotsiantis, and P. Pintelas, 'A novel explainable image classification framework: case study on skin cancer and plant disease prediction', *Neural Comput Appl*, vol. 33, no. 22, pp. 15171–15189, Nov. 2021, doi: 10.1007/S00521-021-06141-0/TABLES/6.

[340] R. Goebel *et al.*, 'Explainable AI: The new 42?', *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11015 LNCS, pp. 295–303, 2018, doi: 10.1007/978-3-319-99740-7_21/FIGURES/6.

[341] A. Abujabal, R. S. Roy, M. Yahya, and G. Weikum, 'QUINT: Interpretable Question Answering over Knowledge Bases', *EMNLP 2017 - Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Proceedings*, pp. 61–66, 2017, doi: 10.18653/V1/D17-2011.

[342] D. Croce, D. Rossini, and R. Basili, 'Auditing Deep Learning processes through Kernel-based Explanatory Models', *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pp. 4037–4046, 2019, doi: 10.18653/V1/D19-1415.

[343] A. Karpathy and L. Fei-Fei, 'Deep Visual-Semantic Alignments for Generating Image Descriptions', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3128–3137.

[344] D. Shin, ‘The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI’, *Int J Hum Comput Stud*, vol. 146, p. 102551, Feb. 2021, doi: 10.1016/J.IJHCS.2020.102551.

[345] D. Shin, B. Zhong, and F. A. Biocca, ‘Beyond user experience: What constitutes algorithmic experiences?’, *Int J Inf Manage*, vol. 52, p. 102061, Jun. 2020, doi: 10.1016/J.IJINFOMGT.2019.102061.

[346] C. Rudin, ‘Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead’, *Nat. Mach. Intell.* 2019. doi: 10.1038/s42256-019-0048-x.

[347] B. Ustun and C. Rudin, ‘Supersparse Linear Integer Models for Optimized Medical Scoring Systems’, *Mach Learn*, vol. 102, no. 3, pp. 349–391, Feb. 2015, doi: 10.1007/s10994-015-5528-6.

[348] Y. Lou, R. Caruana, J. Gehrke, and G. Hooker, ‘Accurate intelligible models with pairwise interactions’, *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, vol. Part F128815, pp. 623–631, Aug. 2013, doi: 10.1145/2487575.2487579.

[349] D. Wei, S. Dash, T. Gao, and O. Günlük, ‘Generalized Linear Rule Models’, *36th International Conference on Machine Learning, ICML 2019*, vol. 2019-June, pp. 11589–11605, Jun. 2019, doi: 10.48550/arxiv.1906.01761.

[350] C. Cortes, M. Mohri, M. Riley, and A. Rostamizadeh, ‘Sample selection bias correction theory’, in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2008. doi: 10.1007/978-3-540-87987-9_8.

[351] C. Winship and R. D. Mare, ‘Models for Sample Selection Bias’, *Annu Rev Sociol*, 1992, doi: 10.1146/annurev.so.18.080192.001551.

[352] S. Wood, ‘CRAN - Package mgcv’. Accessed: Feb. 29, 2024. [Online]. Available: https://cran.r-project.org/web/packages/mgcv/index.html

[353] M. Plummer, ‘Bayesian Graphical Models using MCMC [R package rjags version 4-15]’. Accessed: Feb. 29, 2024. [Online]. Available: https://CRAN.R-project.org/package=rjags

[354] J. Kang, ‘CRAN - Package BayesGPfit’. Accessed: Feb. 29, 2024. [Online]. Available: https://cran.rstudio.com/web/packages/BayesGPfit/index.html

[355] J. F. Bobb, ‘bkmr: Bayesian Kernel Machine Regression’. 2017. [Online]. Available: https://cran.r-project.org/package=bkmr

[356] D. Fink, ‘A Compendium of Conjugate Priors’, 1997, Accessed: Feb. 17, 2023. [Online]. Available: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=089442f59e3f4afb920479a7115c3dc57fb14757

[357] T. Hastie and R. Tibshirani, ‘Generalized additive models: Some applications’, *J Am Stat Assoc*, vol. 82, no. 398, pp. 371–386, 1987, doi: 10.1080/01621459.1987.10478440.

[358] K. Steinnocher, A. de Bono, B. Chatenoux, D. Tiede, and L. Wendt, ‘Estimating urban population patterns from stereo-satellite imagery’,

https://doi.org/10.1080/22797254.2019.1604081, vol. 52, no. sup2, pp. 12–25, Aug. 2019, doi: 10.1080/22797254.2019.1604081.

[359] H. Bagan and Y. Yamagata, ‘Analysis of urban growth and estimating population density using satellite images of nighttime lights and land-use and population data’, *http://dx.doi.org/10.1080/15481603.2015.1072400*, vol. 52, no. 6, pp. 765–780, Nov. 2015, doi: 10.1080/15481603.2015.1072400.

[360] R. A. Houghton *et al.*, ‘Carbon emissions from land use and land-cover change’, *Biogeosciences*, vol. 9, no. 12, pp. 5125–5142, 2012, doi: 10.5194/BG-9-5125-2012.

[361] A. Carpio, R. Ponce-Lopez, and D. F. Lozano-García, ‘Urban form, land use, and cover change and their impact on carbon emissions in the Monterrey Metropolitan area, Mexico’, *Urban Clim*, vol. 39, p. 100947, Sep. 2021, doi: 10.1016/J.UCLIM.2021.100947.

[362] V. Avitabile *et al.*, ‘Carbon emissions from land cover change in Central Vietnam’, *http://dx.doi.org/10.1080/17583004.2016.1254009*, vol. 7, no. 5–6, pp. 333–346, Nov. 2016, doi: 10.1080/17583004.2016.1254009.

[363] S. Angel, S. C. Sheppard, and D. L. Civco, ‘The Dynamics of Global Urban Expansion’, *The World Bank*, 2005.

[364] H. Nuissl and S. Siedentop, ‘Urbanisation and Land Use Change’, 2021. doi: 10.1007/978-3-030-50841-8_5.

[365] P. Olofsson, G. M. Foody, M. Herold, S. v Stehman, C. E. Woodcock, and M. A. Wulder, ‘Good practices for estimating area and assessing accuracy of land change’, *Remote Sens. Environ.*, vol. 148, pp. 42–57, 2014, doi: 10.1016/j.rse.2014.02.015.

[366] P. Olofsson, G. M. Foody, S. v Stehman, and C. E. Woodcock, ‘Making better use of accuracy data in land change studies: Estimating accuracy and area and quantifying uncertainty using stratified estimation’, *Remote Sens. Environ.*, vol. 129, pp. 122–131, 2013, doi: 10.1016/j.rse.2012.10.031.

[367] J. C. Lee and D. J. Sabavala, ‘Bayesian estimation and prediction for the beta-binomial model’, *Journal of Business and Economic Statistics*, vol. 5, no. 3, pp. 357–367, 1987, doi: 10.1080/07350015.1987.10509600.

[368] S. Chib and E. Greenberg, ‘Understanding the metropolis-hastings algorithm’, *American Statistician*, vol. 49, no. 4, pp. 327–335, 1995, doi: 10.1080/00031305.1995.10476177.

[369] J. Burkardt, ‘The Truncated Normal Distribution’, Department of Scientific Computing , Florida State University. Accessed: Feb. 29, 2024. [Online]. Available: https://people.sc.fsu.edu/~jburkardt/presentations/truncated_normal.pdf

[370] T. Amemiya, ‘Tobit models: A survey’, *J Econom*, 1984, doi: 10.1016/0304-4076(84)90074-5.

[371] C. S. Rowland, R. D. Morton, L. Carrasco, G. McShane, A. W. O’Neil, and C. M. Wood, ‘Land Cover Map 2015 (vector, GB)’. NERC Environmental Information Data Centre, 2017. doi: https://doi.org/10.5285/6c6c9203-7333-4d96-88ab-78925e7a4e73.

[372] M. J. Brown *et al.*, ‘Landscape area data 2007 [Countryside Survey].’ NERC Environmental Information Data Centre, 2016. doi: https://doi.org/10.5285/bf189c57-61eb-4339-a7b3-d2e81fdde28d.

[373] L. R. Norton *et al.*, ‘Identifying effective approaches for monitoring national natural capital for policy use’, *Ecosyst Serv*, 2018, doi: 10.1016/j.ecoser.2018.01.017.

[374] J. F. Bobb *et al.*, ‘Bayesian kernel machine regression for estimating the health effects of multi-pollutant mixtures’, *Biostatistics*, 2014, doi: 10.1093/biostatistics/kxu058.

[375] J. Q. Shi, R. Murray-Smith, and D. M. Titterington, ‘Bayesian regression and classification using mixtures of Gaussian processes’, *Int J Adapt Control Signal Process*, 2003, doi: 10.1002/acs.744.

[376] C. F. Dormann *et al.*, ‘Methods to account for spatial autocorrelation in the analysis of species distributional data: A review’, *Ecography*. 2007. doi: 10.1111/j.2007.0906-7590.05171.x.

[377] J. F. Bobb, ‘bkmr: Bayesian Kernel Machine Regression’. 2017.

[378] K. Monteith, J. L. Carroll, K. Seppi, and T. Martinez, ‘Turning Bayesian model averaging into Bayesian model combination’, in *Proc. Int. Jt. Conf. Neural Networks*, 2011. doi: 10.1109/IJCNN.2011.6033566.

[379] A. Berger, T. Gschwantner, R. E. McRoberts, and K. Schadauer, ‘Effects of measurement errors on individual tree stem volume estimates for the Austrian national forest inventory’, *Forest Science*, 2014, doi: 10.5849/forsci.12-164.

[380] J. Breidenbach, C. Anton-Fernandez, H. Petersson, R. E. Mcroberts, and R. Astrup, ‘Quantifying the model-related variability of biomass stock and change estimates in the Norwegian national forest inventory’, *Forest Science*, 2014, doi: 10.5849/forsci.12-137.

[381] R. E. McRoberts, S. V. Stehman, G. C. Liknes, E. Næsset, C. Sannier, and B. F. Walters, ‘The effects of imperfect reference data on remote sensing-assisted estimators of land cover class proportions’, *ISPRS Journal of Photogrammetry and Remote Sensing*, 2018, doi: 10.1016/j.isprsjprs.2018.06.002.

[382] L. See *et al.*, ‘Building a hybrid land cover map with crowdsourcing and geographically weighted regression’, *ISPRS Journal of Photogrammetry and Remote Sensing*, 2015, doi: 10.1016/j.isprsjprs.2014.06.016.

[383] J. C. Laso Bayas *et al.*, ‘A global reference database of crowdsourced cropland data collected using the Geo-Wiki platform’, *Sci Data*, 2017, doi: 10.1038/sdata.2017.136.

[384] S. Fritz *et al.*, ‘Geo-wiki.org: The use of crowdsourcing to improve global land cover’, *Remote Sensing*. 2009. doi: 10.3390/rs1030345.

[385] R. Roy and K. T. George, ‘Detecting insurance claims fraud using machine learning techniques’, *Proceedings of IEEE International Conference on Circuit, Power and Computing Technologies, ICCPCT 2017*, Oct. 2017, doi: 10.1109/ICCPCT.2017.8074258.

[386] Y. Wang and W. Xu, ‘Leveraging deep learning with LDA-based text analytics to detect automobile insurance fraud’, *Decis Support Syst*, vol. 105, pp. 87–95, Jan. 2018, doi: 10.1016/J.DSS.2017.11.001.

[387] B. Itri, Y. Mohamed, Q. Mohammed, and B. Omar, ‘Performance comparative study of machine learning algorithms for automobile insurance fraud detection’, *2019 3rd International Conference on Intelligent Computing in Data Sciences, ICDS 2019*, Oct. 2019, doi: 10.1109/ICDS47004.2019.8942277.

[388] G. Bode, S. Thul, M. Baranski, and D. Müller, 'Real-world application of machine-learning-based fault detection trained with experimental data', *Energy*, vol. 198, p. 117323, May 2020, doi: 10.1016/J.ENERGY.2020.117323.

[389] N. Amruthnath and T. Gupta, 'A research study on unsupervised machine learning algorithms for early fault detection in predictive maintenance', *2018 5th International Conference on Industrial Engineering and Applications, ICIEA 2018*, pp. 355–361, Jun. 2018, doi: 10.1109/IEA.2018.8387124.

[390] C. Yang, J. Liu, Y. Zeng, and G. Xie, 'Real-time condition monitoring and fault detection of components based on machine-learning reconstruction model', *Renew Energy*, vol. 133, pp. 433–441, Apr. 2019, doi: 10.1016/J.RENENE.2018.10.062.

[391] S. Sajjadiani, A. J. Sojourner, J. D. Kammeyer-Mueller, and E. Mykerezi, 'Using Machine Learning to Translate Applicant Work History Into Predictors of Performance and Turnover', *Journal of Applied Psychology*, 2019, doi: 10.1037/APL0000405.

[392] A. A. Mahmoud, T. al Shawabkeh, W. A. Salameh, and I. al Amro, 'Performance Predicting in Hiring Process and Performance Appraisals Using Machine Learning', *2019 10th International Conference on Information and Communication Systems, ICICS 2019*, pp. 110–115, Jun. 2019, doi: 10.1109/IACS.2019.8809154.

[393] M. V. Mariushko, R. E. Pashchenko, and A. S. Nechausov, 'Cloud system ArcGIS online as a managerial decision-making tool in agricultural production', in *Proceedings of 2018 IEEE 9th International Conference on Dependable Systems, Services and Technologies, DESSERT 2018*, 2018. doi: 10.1109/DESSERT.2018.8409190.

[394] M. Cope, E. Mikhailova, C. Post, M. Schlautman, and P. McMillan, 'Developing an integrated cloud-based spatial-temporal system for monitoring phenology', *Ecol Inform*, 2017, doi: 10.1016/j.ecoinf.2017.04.007.

[395] R. Nourjou and M. Hashemipour, 'Smart Energy Utilities based on Real-Time GIS Web Services and Internet of Things', in *Procedia Computer Science*, 2017. doi: 10.1016/j.procs.2017.06.070.

[396] C. De Sousa, L. Fatoyinbo, C. Neigh, F. Boucka, V. Angoue, and T. Larsen, 'Cloud-computing and machine learning in support of country-level land cover and ecosystem extent mapping in Liberia and Gabon', *PLoS One*, 2020, doi: 10.1371/journal.pone.0227438.

[397] D. Marković, A. Mizrahi, D. Querlioz, and J. Grollier, 'Physics for neuromorphic computing', *Nature Reviews Physics 2020 2:9*, vol. 2, no. 9, pp. 499–510, Jul. 2020, doi: 10.1038/s42254-020-0208-2.

[398] L. Chen, T.-Y. Wang, S.-J. Ding, and S. Furber, 'Large-scale neuromorphic computing systems', *J Neural Eng*, vol. 13, no. 5, p. 051001, Aug. 2016, doi: 10.1088/1741-2560/13/5/051001.

[399] A. Lye, A. Cicirello, and E. Patelli, 'Sampling methods for solving Bayesian model updating problems: A tutorial', *Mech Syst Signal Process*, vol. 159, p. 107760, Oct. 2021, doi: 10.1016/J.YMSSP.2021.107760.

[400] J. Bierkens, P. Fearnhead, and G. Roberts, 'The Zig-Zag process and super-efficient sampling for Bayesian analysis of big data', *https://doi.org/10.1214/18-AOS1715*, vol. 47, no. 3, pp. 1288–1320, Jun. 2019, doi: 10.1214/18-AOS1715.

[401] F. Feroz and M. P. Hobson, ‘Multimodal nested sampling: An efficient and robust alternative to Markov Chain Monte Carlo methods for astronomical data analyses’, *Mon Not R Astron Soc*, vol. 384, no. 2, pp. 449–463, Feb. 2008, doi: 10.1111/J.1365-2966.2007.12353.X/2/M_MNRAS0384-0449-MU45.GIF.

[402] P. I. Frazier, ‘A Tutorial on Bayesian Optimization’, Jul. 2018, Accessed: Feb. 27, 2024. [Online]. Available: https://arxiv.org/abs/1807.02811v1

[403] E. G. Ryan, C. C. Drovandi, J. M. Mcgree, and A. N. Pettitt, ‘A Review of Modern Computational Algorithms for Bayesian Optimal Design’, *International Statistical Review*, vol. 84, no. 1, pp. 128–154, Apr. 2016, doi: 10.1111/INSR.12107.