A probabilistic approach to estimating timber harvest location

Jakub Truszkowski[1,2,†], Roi Maor[3], Raquib Bin Yousuf[4], Subhodip Biswas[4], Caspar Chater[3], Peter Gasson[3], Scot McQueen[5], Marigold Norman[6], Jade Saunders[6], John Simeone[7], Naren Ramakrishnan[4], Alexandre Antonelli[1,2,3,8,*], and Victor Deklerck[3,*]

[1]Department of Biological and Environmental Sciences, University of Gothenburg, Gothenburg, Sweden
[2]Gothenburg Global Biodiversity Centre, Gothenburg, Sweden
[3]Royal Botanic Gardens, Kew, Richmond, UK
[4]Department of Computer Science, Virginia Tech, Arlington, Virginia 22203, USA
[5]Forest Stewardship Council International, Technology and Information Unit, Bonn, Germany
[6]World Forest ID, Thomas Circle NW, Suite 700, Washington DC 20005, USA
[7]Simeone Consulting, LLC, Littleton, New Hampshire, USA
[8]Department of Plant Sciences, University of Oxford, Oxford, UK
[*]Co-senior authors
[†]Corresponding author: jakub.truszkowski@bioenv.gu.se, tel: +46734782603

November 22, 2024

Abstract

Determining the harvest location of timber is crucial to enforcing international regulations designed to protect natural resources and to tackle illegal logging and associated trade in forest products. Stable Isotope Ratio Analysis (SIRA) can be used to verify claims of timber harvest location by matching levels of naturally-occurring stable isotopes within wood tissue to location-specific ratios predicted from reference data ('isoscapes'). However, overly simple models for predicting isoscapes have so far limited the confidence in derived estimates of timber provenance. In addition, most use cases have limited themselves to differentiating between a small number of pre-determined location options. Here, we present a new SIRA data analysis pipeline, designed to infer the harvest location of a focal tree out of a continuous, arbitrarily large area. We use Gaussian Processes to robustly estimate isoscapes from reference wood samples, and overlay with species distribution data to compute, for every pixel in the study area, the probability of it being the origin of the examined timber. This is the first time, to our knowledge, that this approach is applied to determining timber provenance, providing probabilistic results rather than a binary outcome. Additionally, we include an active learning tool to identify locations from which additional samples would maximize the improvement to model performance, allowing for optimisation of field efforts. We demonstrate our approach on a set of SIRA data from seven oak species in the USA as a proof of concept. Our method can determine the harvest location up to within 520 km from the true origin of the sample and outperforms the state-of-the-art approach. Incorporating species distribution data improves accuracy by up to 36%. The future-sampling locations proposed by our tool decrease the variance of resultant isoscapes by up to 86% more than sampling the same number of locations at random. The method we present here greatly advances the toolset available for verification of

timber harvest location, will empower authorities worldwide in enforcing anti-deforestation legislation and will help protect natural resources.

Keywords: stable isotopes, origin traceability, timber provenance, illegal logging, isoscapes, Gaussian Processes

1 Introduction

Unsustainable exploitation of natural resources is the largest driver of terrestrial biodiversity loss after land-use change (Díaz et al., 2019) and a major conservation challenge globally. To prevent a sixth mass extinction (Barnosky et al., 2011), nearly 200 nations have recently agreed on a new set of targets and goals under the Kunming-Montreal Global Biodiversity Framework. In particular, Target 5 of the agreement includes the objective to "ensure that the use, harvesting and trade of wild species is sustainable, safe and legal, preventing overexploitation" (2022 UN Biodiversity Conference, 2022). Meeting this ambitious target will require overcoming a key element of unsustainable use of natural resources: the illegal harvest of threatened tree species.

Legal frameworks have been established to combat illegal logging and trade in illegally harvested timber, such as the Convention on the International Trade in Endangered Species (CITES), the US Lacey Act (amended 2008), the UK Timber Regulation (2021), the EU Deforestation Regulation (EUDR; 2023) and the Australian Illegal Logging Prohibition Act (2012). The new policies place additional traceability and reporting requirements on companies trading in wood and agricultural products (Dormontt et al., 2015). For example, the EUDR requires operators to record and report the coordinates of production location (forest or farm), and enforcement officials will be expected to scrutinize those claims of harvest location. Despite the

comprehensive legislation already in place and the international commitments under current adoption, enforcement of such regulations remains a challenge. Illegally harvested timber is shipped under false declarations of origin or mixed into legal shipments, and methods for verifying geographical location have so far only been able to determine the correct location out of a few pre-determined options, mostly at country-level resolution (Watkinson et al., 2022a; Horacek et al., 2009; Muñoz-Redondo et al., 2021). This challenge is greatly intensified by the new EUDR legislation adopting precise geographical location (GPS point or polygon for plot of land) as a determinant of the legal status of timber.

1.1 Stable Isotope Ratio Analysis to verify provenance

Well-established scientific techniques enable measurement of the chemical, anatomical and genetic features of plants from a tissue sample (Deklerck, 2023), with ever increasing precision and availability. When compared against a robust physical reference collection, these attributes of the tissue can be used to (in-)validate declared species and origin claims, and support enforcement officials in their efforts to detect, for example, illegally harvested timber or fraud in supply chains.

Stable isotope ratio analysis (SIRA) is one of the most promising technologies in this context. Several elements within biological tissues (mainly Hydrogen, Oxygen, Carbon, Sulfur, Nitrogen) have multiple naturally-occurring stable isotopes, whose ratios vary predictably across space, in correlation with environmental conditions (West et al., 2010; Siegwolf et al., 2022; Gay et al., 2022; Pederzani and Britton, 2019). The heavy isotopes of these elements do not undergo radioactive decay, and their proportion can be readily detected by mass spectrometry (Boner et al., 2007). The isotopic composition of elements incorporated into the tissues of a plant is determined by soil properties, climate, metabolic fractionation and other biotic and abiotic

conditions characteristic of the species and the habitat in which the individual grows (Siegwolf et al., 2022; Camin et al., 2017; Horacek et al., 2009; van der Sleen et al., 2017; Gay et al., 2022). Hence, differences in stable isotope ratios among individuals correspond to the environment they grew in, and can be used to discriminate between plants from different geographic areas.SIRA has proven useful in determining risk of illegally harvested material in a wide variety of contexts, for example, forest products (Watkinson et al., 2020; Boner et al., 2007), wildlife trafficking (Bowen et al., 2005; Koehler et al., 2019; Wunder and Norris, 2008; Vander Zanden et al., 2015), ivory trade (Van der Merwe et al., 1990; Ziegler et al., 2016), agricultural products (Camin et al., 2016; Saadat et al., 2022), fish/seafood (Cusa et al., 2022; Silva et al., 2021; Kroetz et al., 2020), precious metals (Kirk et al., 2003), and natural and synthetic illegal drugs (Kurashima et al., 2004; Casale et al., 2005), but without the spatial discrimination required by the new timber legislation.

1.2 Modelling approach

Current modelling practices for the use of SIRA to verify harvest location of both legally and illegally harvested forest products require improvement. The use of SIRA is currently limited by the simplistic models used, as well as by the limited number of reference samples used as input data for such models. Reference sampling campaigns are costly and budgetary needs are often underestimated, with sampling locations often taking into account relative ease of sampling rather than areas that yield a gain in model prediction accuracy (Schmitz et al., 2019). There has been considerable development of isoscapes ("isotope landscapes"), given that stable isotopic variation is a continuous spatial variable in nature (West et al., 2010). These isoscapes are geospatial maps that show the isotopic variation of the material of interest (West et al., 2010). While the potential of isoscapes for determining forest product origins has long been

recognized, few rigorous methods exist to achieve this task. The existing methods use simple prediction strategies such as linear regression (Watkinson et al., 2020, 2022b), which do not fully leverage the information contained in isotope ratio data.

Gaussian Process (GP) regression, which is closely related to kriging in geostatistics literature, is a class of flexible regression models which use the values in sampled points to estimate the values in surrounding points (Li and Heap, 2008; Deklerck, 2023; Williams and Rasmussen, 2006). A key advantage of GP regression is that it can quantify the uncertainty of its own predictions based on the inferred spatial covariance structure of the samples. The importance of quantifying the uncertainty of predictions is increasingly recognized in safety-critical (Jankowiak et al., 2020) and forensic (Su and Srihari, 2010; Swofford and Champod, 2022) machine learning applications. Additionally, GP regression facilitates inference of a sparsely sampled variable of interest from variables that are highly correlated with it but more densely sampled (Adhikary et al., 2017; Kanankege et al., 2018). In the context of plant harvest location estimation, this translates to inferring stable isotope ratios from atmospheric drivers (such as precipitation, temperature and water vapor pressure) known to influence the stable isotope signal in wood (Horacek et al., 2009; Siegwolf et al., 2022). This then provides a powerful tool for predicting the isotopic composition in areas that have not yet been sampled. However, previous work on timber isoscapes used GP regression primarily as a spatial interpolation technique without a probabilistic interpretation (Gori et al., 2018; Watkinson et al., 2022a). Others used approximate GP models to derive variance estimates for origin determination in animals (Ma et al., 2020; St. John Glew et al., 2019).

Here, we develop GP-based probabilistic machine learning models to infer timber harvest location by directly modelling timber isoscapes from SIRA data, with the aid of atmospheric predictors and species distribution data. We show that probabilistic modeling greatly enhances

the utility of SIRA in estimating the geographical origin of timber, and, assisted by a reference dataset (Gasson et al., 2021), can be used to guide future sample collection by identifying sampling locations that will minimize prediction uncertainty.

2 Materials and Methods

2.1 Data sets

We use data from 87 trees of the genus *Quercus*, sampled across the contiguous United States, as described in (Watkinson et al., 2020). Stable isotope ratio measurements were done following the protocol described in (Boner et al., 2007). Each entry contained stable isotope ratio measurements of oxygen δ^{18}O (ratio between ^{18}O and ^{16}O), hydrogen δ^{2}H (ratio between ^{2}H and ^{1}H), carbon δ^{13}C (ratio between ^{13}C and ^{12}C) and sulfur δ^{34}S (ratio between ^{34}S and ^{32}S) as well as the GPS coordinates of the sampled tree. As stable isotope ratios are largely driven by environmental conditions such as precipitation, temperature, humidity and so on, publicly available datasets for these factors are used to aid the inference of isoscapes. We used the following atmospheric data: δ^{2}H and δ^{18}O isotopic composition of precipitation (Bowen and Revenaugh, 2003), water vapor (Borbas, 2015) (found to be associated with δ^{13}C by (Watkinson et al., 2020)), reflected shortwave radiation (NEO, 2023) and precipitation (multi-satellite) (Huffman et al., 2020), both of which were found to be associated with δ^{34}S (Watkinson et al., 2020). For each of those data types, we used monthly means averaged over a number of years to minimize the impact of weather patterns in specific years (see (Watkinson et al., 2020) for precise year ranges).

To inform the priors (probability distributions representing the prior belief on possible tree locations) of the models we develop, we used species inventory data across the natural range of

each species within the United States (Wilson et al., 2013), downloaded from: https://www.fs.usda.gov/rds/archive/Catalog/RDS-2013-0013 on 09/12/2022. This data is available as species-specific raster layers of tree abundance at 250m resolution. We then used the function project() of the R package $terra$ (Hijmans, 2022) to bilinearly aggregate abundance data so that it matched the spatial resolution of other spatial data in the pipeline.

2.2 Model architecture

Figure 1 presents an overview of the data sets and components comprising our model and output. We use a rectangular grid to represent the study area. Grid points are placed every 0.125 degree latitude (≈ 14 km) and every 0.06 degree longitude (≈ 4.3-6.0 km), which allows us to approximate spatial probability distributions with high accuracy. For every isotope ratio (IR), we fit a GP regression model to the training data to obtain the posterior mean and variance of the isotope ratio for every point of the grid (see Appendix for the full detail on implementation).

Gaussian Processes are a class of flexible regression models, which enable the modeler to quantify the uncertainty about specific predictions. A GP regression model assumes that the responses (in our case, isotope ratios) at different locations are jointly normally distributed (Gaussian). The model is defined by three elements: (1) the mean, for which we use a constant, (2) the covariance function for which we use the Matern function (Williams and Rasmussen, 2006) with separate scaling parameters for latitude and longitude and (3) the noise parameter. The choice of mean and covariance functions reflects prior knowledge and modelling assumptions about the regression problem. The covariance function expresses the amount of information about unobserved locations contained in nearby observed values. The function parameters as well as the noise parameter are estimated by maximizing the likelihood of the

training data, in contrast to standard kriging approaches in geostatistics literature, which use approximate techniques based on summary statistics. We use GPyTorch (Gardner et al., 2018) to efficiently find the maximum likelihood parameter estimates.

The input to the GP consists of the coordinates and/or the climate variable values at the grid point. For a combination of observed stable isotope ratios (y_O, y_H, y_C, y_S) (meaning δ^{18}O, δ^{2}H, δ^{13}C, δ^{34}S), we compute the likelihood of this observation at every point in the grid, using the four GP regression models estimated in the previous step. This likelihood is the product of likelihoods for each isotope ratio as we assume independence between isotopes. Given the prior and the likelihood, we compute the posterior probability of each grid point being the harvest location of the sample by multiplying the prior and the likelihood and normalizing so that the probabilities sum up to 1. For ease of interpretation, the output is a map with highest-posterior density (HPD) regions indicated for several probability levels (15%, 30%, 50%, 75%, 90%, 95%).

To incorporate atmospheric data into the isoscape we use monthly averages of the atmospheric variables listed in Section 2.1. We use a linear covariance term to model the covariance component corresponding to the variation in the respective atmospheric variables. The linear covariance function models a linear relationship between the atmospheric variable and the response and is mathematically equivalent to Bayesian linear regression (Williams and Rasmussen, 2006). The overall covariance function is then the sum of the spatial and linear terms and can be seen in Appendix A.

We use the spatial density maps developed by (Wilson et al., 2013) to design two prior distributions for sample origin that account for the spatial distribution of oak species. The first, which we call the *density prior*, holds that the probability of a sample originating at a grid cell is proportional to the basal area (average amount of area occupied by tree stems per unit of

space) recorded at the grid cell. The second, which we call the *range prior*, assigns equal probability to every grid cell where above-zero basal area has been recorded. In addition, both priors allow for a small probability that a sample might occur outside its observed range - we set that probability to 0.01 and diffuse it uniformly over all grid points within the contiguous United States where the species does not occur according to (Wilson et al., 2013).

2.3 Performance evaluation

We perform 5-fold cross-validation on the data set and report the average values of all performance metrics over all data points. Samples with incomplete or ambiguous species information and samples collected in botanical gardens outside of their species' native range are excluded from the test sets, but not from the training sets, resulting in a total of 74 test samples across the 5 folds. We report performance of our models as well as our implementation of the approach by (Watkinson et al., 2020) averaged across the five cross-validation folds. Rigorously evaluating the performance of our models is a non-trivial task as each model produces a distribution over possible locations, rather than a single location. For this reason, we have defined several metrics to investigate different aspects of probabilistic harvest location prediction:

1. Predictive log-likelihood and log-posterior: We report the log-likelihood and the log-posterior of observing the sample at its true origin. Both of those measure how well the model fits the test data.

2. Mode distance: We report the great circle distance between the true location and the *mode* of the posterior distribution, i.e. the highest scored location according to the model. This metric measures the accuracy of the highest-scored locations, but it does not

account for the amount of uncertainty in model predictions.

3. Mean absolute error (MAE): To investigate how distant our predicted locations are from the truth, we report the expected distance between the true location and a location sampled randomly from the posterior distribution returned by our model

$$MAE = \int_{\mathbf{x} \in A} d(\mathbf{x}_t, \mathbf{x}) p(\mathbf{x}|\bar{y}, S) d\mathbf{x}$$

where $d()$ is the great circle distance between the two points and $p(\mathbf{x}|\bar{y}, S)$ is the posterior probability of $\mathbf{x}$ being the harvest location. A perfect prediction would have the distance of 0. This metric will favour predictions concentrated around the true location over equally dispersed predictions concentrated elsewhere. It will also favour less dispersed predictions generally. For the method of Watkinson *et al.*, which only outputs a region of plausible locations, we assume a uniform distribution within the region highlighted by the model. In practice, isoscapes often predict similar isotope ratio values at distant locations, so even a statistically efficient method might yield a high MAE value.

4. Area scored higher than the true location (ASH): The behaviour of MAE is influenced by the shape of the posterior distribution, which favours unimodal over multimodal shapes. We report the total surface area corresponding to the points that the model considers more plausible than the true origin of the sample.

$$ASH = \int_{\mathbf{x} \in A} I\left[score(\mathbf{x}|\bar{y}, S) > score(\mathbf{x}_t|\bar{y}, S)\right] d\mathbf{x}$$

where $I(.)$ is the indicator function that yields 1 when the statement is true and 0 otherwise. For all GP models, the score is the posterior probability of harvest location,

whereas for the method of Watkinson *et al.* we take the score to be the negative of the minimum value of the threshold that results in the true location being included in the highlighted region. In contrast to MAE, this metric is likely to give a low value to a posterior distribution that is concentrated in several small areas as long as one of those areas contains the true location. For example, if the true location could be a county in New York or a county in West Virginia, this would give a low ASH but high MAE as the two counties are far apart.

2.4 Guiding future sampling efforts

Field sample collections are time-consuming and expensive. We can optimize future field collections using informed prediction of where additional sample points are most needed for increasing origin estimation accuracy. The isoscape variance estimates provided by GPs can be used to guide future sampling efforts, which in turn will maximize the performance of the model. This approach is known as *active learning* in the machine learning literature. Here, we propose a strategy to minimize the error of our isoscape estimates by carefully choosing future sampling locations.

Early attempts at efficient active learning in GPs involved collecting samples at points with highest response variance or, equivalently, picking a set of points that maximizes the entropy of responses (Cressie, 2015). Unfortunately, this approach tends to recommend collecting samples on the boundaries of the study area, which is wasteful as the newly collected samples improve isoscapes in a smaller fraction of the study area than if they were placed away from the boundary. This motivated researchers to propose several criteria for optimizing sampling (Guestrin et al., 2005; Ramakrishnan et al., 2005). Here, we adopt an approach similar to that of (Guestrin et al., 2005) with a few modifications designed to address the large size of

our spatial grid, which renders their original method computationally intractable for our data set. We seek to maximize the *average* reduction in predictive variance across our study area that can be achieved by adding a sample to the training set. With S the set of sampled locations and G the set of grid points, we define the information gain (IG) associated with adding a new point $(\mathbf{x}^*)$ to the training data set as

$$IG(\mathbf{x}^*) = \sum_{\mathbf{x} \in G} \left[(\log(\sigma^2(\mathbf{x}|S)) - \log(\sigma^2(\mathbf{x}|S \cup \{\mathbf{x}^*\})) \right] \tag{1}$$

where the predictive variances are computed using Equation A.4. The algorithm then picks the point in the grid that yields the highest IG. Importantly, the predictive variances depend only on the sampling locations and not on the sampled values, so it is possible to sequentially propose multiple sampling points before collecting the samples. Our method sequentially proposes sample collection points until a user-specified number of samples is reached. We assume that samples can only be collected in locations where at least one of the species is present. Thus, grid points that lie outside every species range are excluded from the procedure. Our active learning strategy requires repeatedly computing a large number of predictive variances for varying training sets. To reduce computation time, we randomly downsample our grid to 15000 points before running the analysis. In addition, we assume that the reduction in variance is negligible for grid points situated more than 15 degrees away from the newly sampled point $(\mathbf{x}^*)$ in longitude or more than 7.5 degrees in latitude.

3 Results

3.1 Model accuracy and comparison

The plausible location areas identified by our models consisted of points within an average distance of 520-870 kilometers from the true location of the oak tree samples, depending on model settings. Even with a relatively small training data set of $69 - 70$ training samples (depending on the cross-validation fold), our model is able to exclude the vast majority of the study area from consideration as a possible source of the focal sample. All our models outperform the state-of-the-art method for determining timber harvest location (Watkinson et al., 2020) in most or all metrics. Table 1 shows performance metrics for all the models on the test data set.

Incorporating species distribution information improves prediction performance for every model and every metric examined except the log-likelihood, which is computed independently of the prior. Informative priors improve MAE by 16% to 35% and ASH by 15% to 57% with most improvement for the pure spatial model and least for the spatial+atmospheric model. The more informative density prior gives better accuracy than the range prior according to all the metrics. Predicted probability maps for a few test points are shown in Fig. 2 and 3.

The spatial-only GP model gives the closest location predictions to the true location of the tree samples, except when a flat prior is used. In general, the spatial-only model and the combined spatial+atmospheric model give similar results on all metrics and outperform the atmospheric-only model in almost all settings. Somewhat surprisingly, the combined model does not outperform the spatial-only model. This might be due to the relatively small dataset used here or the choice of atmospheric predictors, and remains to be tested as we continue to expand our reference databases. The predictions of atmospheric GP models appear qualitatively

different from those from the purely spatial GP, perhaps because atmospheric model predictions emphasize geographical areas with distinct climate patterns, such as Appalachia or the Gulf Coast. Unsurprisingly, the purely spatial GP identifies areas that are more spatially cohesive but do not share any obvious physical features.

3.2 Guiding future sampling efforts

We investigated the performance of our active learning strategy on the US oak data set. For the spatial-only model, we let our method propose 10 new sampling locations to add to the training data set in the first cross-validation fold and computed the predictive variances before and after including the proposed locations.

The resulting isoscape standard deviation maps are shown in Figure 4. Our active learning strategy proposes sampling locations in currently undersampled regions with high predictive variance and sampling in those areas results in a visible improvement. The highest decrease in predictive variance was observed for δ^2H while the lowest decrease was observed for δ^{18}C. Most of the chosen locations are close to, but not at the boundary of, the allowed sampling area.

To investigate the efficiency of our active learning procedure, we compared isoscape variances resulting from active learning with those resulting from adding the same number of points sampled randomly from the allowed sampling area. We generated 100 such variance maps and compared the average variance (across the allowed sampling area) of those maps with the maps in Fig. 4. Appendix B shows the average predictive variances as a function of the number of points added for both random and active learning sampling strategies. We see that our active learning strategy results in a substantially faster decrease in predictive variances. After adding 10 samples, the reduction in variance achieved by our active learning method is between 64% ($\delta^{13}C$) and 86% ($\delta^{18}O$) greater than the average reduction achieved by the same number of

random samples.

4 Discussion

4.1 Harvest location estimation

To halt illegal logging, to enforce timber regulations and to protect biodiversity in forested landscapes, we need to be able to accurately estimate timber harvest location. Although several examples exist of applying SIRA for timber origin questions (Gori et al., 2018; Watkinson et al., 2020; Kagawa and Leavitt, 2010), these approaches do not take full advantage of (1) atmospheric and species distribution datasets available or (2) state-of-the-art probabilistic machine learning models. In addition, many SIRA use-cases limit themselves to a classification problem (country X versus country Y) compared to a continuous assignment problem (true harvest location). In response to growing evidence of fraud in supply chains, legislation increasingly requires operators to trace back to plot (for example the EU Deforestation Regulation). Consequently, determining the true harvest location will likely become increasingly important. In this work we present a new computational pipeline which aims at taking advantage of both (1) and (2) while predicting the harvest location as a continuous variable. The accuracy of our models depends on the specific modelling approach and the data sets used. Using prior information about species distribution results in a considerable increase in accuracy regardless of which model is used by all metrics considered. The impact of adding species distribution data appears to be greater for the spatial-only model than models that use atmospheric information. This could be due to climate patterns influencing both species distributions (habitat suitability) and the values of the atmospheric variables that we incorporated in our models, which renders species distribution information more redundant once

atmospheric variables have been included in the model.

Within timber tracing literature, our method bears the most resemblance to the work of (Watkinson et al., 2020), which uses linear regression to predict isoscapes based on atmospheric data. Their approach assumes a constant variance across the study area. In contrast, our method estimates the predictive variances based on the spatial covariance structure learned from the reference data, which enables us to translate differences in sampling density across regions into varying levels of confidence in isoscapes across space. Our method also assumes a linear relationship between atmospheric predictors and isoscapes, but our GP formulation implicitly integrates over plausible values of regression parameters, which should lead to more robust predictions compared to standard linear regression. In addition, our approach makes use of species distribution data, which yields substantially improved predictions compared to uninformative priors. Finally, our approach enables us to propose locations for further sample collection that maximize the utility of the samples.

The estimation of spatial covariance structure has recently attracted attention in animal stable isotope studies. (Ma et al., 2020) recently proposed a method that uses probabilistic precipitation isoscapes derived from a GP (Courtiol et al., 2019), which are then calibrated to produce isoscapes for the species of interest. (St. John Glew et al., 2019) introduced a model combining spatial and environmental effects using a novel likelihood approximation for isoscape estimation, though the main focus of their work is isoscape modelling, not origin estimation. These approaches differ from ours in that 1) they rely on Laplace approximations for isoscape estimation rather than exact likelihood maximization; 2) they use ordinary least-squares regression to account for atmospheric predictors, whereas our method uses a Bayesian approach via a linear covariance term; and 3) they do not aim to actively improve isoscapes through additional sampling. A common feature between these models and ours is using a grid to

compute the posterior distribution of origins, which to the best of our knowledge was first considered by (Wunder, 2010).

Our current best performing model can estimate the harvest location for *Quercus* species to 520 km across the east of the United States. Future field expeditions will lead to an improvement, especially if the identified priority locations are targeted (see 4.2). The presented model will be adapted to other use cases, with a focus largely on endangered tropical species which are under high logging pressure.

We expect that our models will be more accurate once more timber samples become available. The size of the current data set of wood samples available to this study (87 samples) is quite small relative to the area of the contiguous United States, which inevitably results in large predictive variance in many areas. In addition to reducing uncertainty about undersampled areas, larger data sets (in the range of hundreds to thousands of samples collected from across the US) should also enable researchers to use more complex GP models, including models with heterogeneous noise (Binois et al., 2018), or deep GP models where the covariance function is modelled by a neural network (Wilson et al., 2016).

4.2 Guiding future collection efforts

Under the World Forest ID Programme (Gasson et al., 2021), tens of thousands of tree samples are being collected globally, and are being analysed by different techniques, including SIRA, to build georeferenced databases which can be used to identify timber harvest origin. Our active learning approach can be used to inform future sample collection efforts and increase model accuracy that can be achieved within a fixed sampling budget. This will be especially important in tropical regions, where reaching sampling sites can be difficult, time intensive and expensive. A good sampling design can substantially improve model performance (Contina et al., 2022),

and our method can be used to adapt sampling efforts as more data is analysed. Our current approach focuses on minimizing predictive variances without considering the impact of newly sampled points on model parameters. Extending our approach to *non-myopic* sampling (Krause and Guestrin, 2007), which considers the impact on model parameters, would constitute an interesting future research direction. Another avenue for improving our approach would be to augment our IG criterion to reflect the varying investment in collecting samples as a function of the time, logistics, and financial cost of reaching the desired sampling location.

5 Conclusion

The accurate estimation of geographic origin of globally traded wood products is a critical step in combating illegal logging and associated trade, by supporting authorities' ability to verify claims made by traders at any supply chain node. In this work we presented a novel analytical pipeline that brings together and incorporates multiple data types and algorithms. This methodology is able to accurately predict timber product origin and can be used to optimize future field sampling to further increase accuracy and precision. We hope that this work will inspire more efforts to expand reference collections of wood samples, and that governments and companies will more routinely use the technological tools at their disposal to have more oversight over their supply chains and promote a more sustainable use of natural resources.

6 Acknowledgements

Jakub Truszkowski and Alexandre Antonelli are funded by the Swedish Research Council (grant number 2019-05191). Victor Deklerck is funded under the World Forest ID Timber at Kew

Grant provided by the Department of Environment, Food & Rural Affairs (DEFRA), International Climate Finance (ICF) R&D Programme, UK (project 29084). Caspar Chater and Roi Maor are funded under the World Forest ID FRC at Kew grant provided by DEFRA, ICF R&D Programme, UK. Alexandre Antonelli also acknowledges financial support from the Swedish Foundation for Strategic Environmental Research MISTRA (Project BioPath) and the Royal Botanic Gardens, Kew. The authors want to thank the US Forest Service - International Programs and FSC-US for the initial collection of the US dataset. The work of Norman, Saunders, Simeone, and Ramakrishnan was supported in part by US National Science Foundation grant CMMI-2240402. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF.

7 Conflict of interest statement

The authors declare that they have no conflicts of interest.

References

2022 UN Biodiversity Conference. The "kunming-montreal global biodiversity framework", 2022. URL `https://www.cbd.int/gbf/targets/`.

Sajal Kumar Adhikary, Nitin Muttil, and Abdullah Gokhan Yilmaz. Cokriging for enhanced spatial interpolation of rainfall in two australian catchments. *Hydrological processes*, 31(12): 2143–2161, 2017.

Anthony D Barnosky, Nicholas Matzke, Susumu Tomiya, Guinevere OU Wogan, Brian Swartz,

Tiago B Quental, Charles Marshall, Jenny L McGuire, Emily L Lindsey, Kaitlin C Maguire, et al. Has the earth's sixth mass extinction already arrived? *Nature*, 471(7336):51–57, 2011.

Mickael Binois, Robert B Gramacy, and Mike Ludkovski. Practical heteroscedastic gaussian process modeling for large simulation experiments. *Journal of Computational and Graphical Statistics*, 27(4):808–821, 2018.

M Boner, Th Sommer, C Erven, and Hilmar Förstel. Stable isotopes as a tool to trace back the origin of wood. In *Proceedings of the international workshop, Fingerprinting methods for the identification of timber origins, October*, pages 8–9, 2007.

et al. Borbas, E. Terra/modis temperature and water vapor profiles 5-min l2 swath 5km, 2015. URL `http://dx.doi.org/10.5067/MODIS/MOD07_L2.061`.

Gabriel J Bowen and Justin Revenaugh. Interpolating the isotopic composition of modern meteoric precipitation. *Water resources research*, 39(10), 2003.

Gabriel J Bowen, Leonard I Wassenaar, and Keith A Hobson. Global application of stable hydrogen and oxygen isotopes to wildlife forensics. *Oecologia*, 143(3):337–348, 2005.

Federica Camin, Luana Bontempo, Matteo Perini, and Edi Piasentier. Stable isotope ratio analysis for assessing the authenticity of food of animal origin. *Comprehensive Reviews in Food Science and Food Safety*, 15(5):868–877, 2016.

Federica Camin, Markus Boner, Luana Bontempo, Carsten Fauhl-Hassek, Simon D. Kelly, Janet Riedl, and Andreas Rossmann. Stable isotope techniques for verifying the declared geographical origin of food in legal cases. *Trends in Food Science & Technology*, 61:176–187, 2017. ISSN 0924-2244.

John F Casale, James R Ehleringer, David R Morello, and Michael J Lott. Isotopic fractionation of carbon and nitrogen during the illicit processing of cocaine and heroin in south america. *Journal of Forensic Science*, 50(6):JFS2005077–7, 2005.

Andrea Contina, Sarah Magozzi, Hannah B Vander Zanden, Gabriel J Bowen, and Michael B Wunder. Optimizing stable isotope sampling design in terrestrial movement ecology research. *Methods in Ecology and Evolution*, 13(6):1237–1249, 2022.

Alexandre Courtiol, François Rousset, Marie-Sophie Rohwäder, David X Soto, Linn S Lehnert, Christian C Voigt, Keith A Hobson, Leonard I Wassenaar, and Stephanie Kramer-Schadt. Isoscape computation and inference of spatial origins with mixed models using the r package isorix. In *Tracking animal migration with stable isotopes*, pages 207–236. Elsevier, 2019.

Noel Cressie. *Statistics for spatial data*. John Wiley & Sons, 2015.

Marine Cusa, Katie St John Glew, Clive Trueman, Stefano Mariani, Leah Buckley, Francis Neat, and Catherine Longo. A future for seafood point-of-origin testing using DNA and stable isotope signatures. *Reviews in Fish Biology and Fisheries*, 32(2):597–621, June 2022. ISSN 0960-3166, 1573-5184. doi: 10.1007/s11160-021-09680-w.

V. Deklerck. Timber origin verification using mass spectrometry: Challenges, opportunities, and way forward. *Forensic Science International: Animals and Environments*, 3:100057, 2023. ISSN 2666-9374.

Sandra Díaz, Josef Settele, Eduardo S Brondízio, Hien T Ngo, John Agard, Almut Arneth, Patricia Balvanera, Kate A Brauman, Stuart HM Butchart, Kai MA Chan, et al. Pervasive human-driven decline of life on earth points to the need for transformative change. *Science*, 366(6471):eaax3100, 2019.

Eleanor E Dormontt, Markus Boner, Birgit Braun, Gerhard Breulmann, Bernd Degen, Edgard Espinoza, Shelley Gardner, Phil Guillery, John C Hermanson, Gerald Koch, et al. Forensic timber identification: It's time to integrate disciplines to combat illegal logging. *Biological Conservation*, 191:790–798, 2015.

Jacob R Gardner, Geoff Pleiss, David Bindel, Kilian Q Weinberger, and Andrew Gordon Wilson. Gpytorch: Blackbox matrix-matrix gaussian process inference with gpu acceleration. In *Advances in Neural Information Processing Systems*, 2018.

Peter E Gasson, Cady A Lancaster, Roger Young, Sara Redstone, Isabella A Miles-Bunch, Gareth Rees, R Philip Guillery, Meaghan Parker-Forney, and Elizabeth T Lebow. Worldforestid: Addressing the need for standardized wood reference collections to support authentication analysis technologies; a way forward for checking the origin and identity of traded timber. *Plants, People, Planet*, 3(2):130–141, 2021.

Justin D. Gay, Bryce Currey, and E. N.J. Brookshire. Global distribution and climate sensitivity of the tropical montane forest nitrogen cycle. *Nature Communications*, 13, 12 2022. ISSN 20411723. doi: $10.1038/s41467\text{-}022\text{-}35170\text{-}z$.

Yuri Gori, Ana Stradiotti, and Federica Camin. Timber isoscapes. a case study in a mountain area in the italian alps. *PLoS One*, 13(2):e0192970, 2018.

Carlos Guestrin, Andreas Krause, and Ajit Paul Singh. Near-optimal sensor placements in gaussian processes. In *Proceedings of the 22nd international conference on Machine learning*, pages 265–272, 2005.

Robert J. Hijmans. terra: Spatial data analysis, 2022. URL `https://CRAN.R-project.org/package=terra`. R package version 1.6-17.

Micha Horacek, Michael Jakusch, and Hannes Krehan. Control of origin of larch wood: discrimination between european (austrian) and siberian origin by stable isotope analysis. *Rapid Communications in Mass Spectrometry*, 23:3688–3692, 2009. ISSN 0924-2244.

G.J. Huffman, A. Behrangi, D.T. Bolvin, and E.J. Nelkin. Gpcp version 3.1 satellite-gauge (sg) combined precipitation data set, 2020. URL `https://disc.gsfc.nasa.gov/datasets/GPCPMON_3.1/summary`.

Martin Jankowiak, Geoff Pleiss, and Jacob Gardner. Parametric gaussian process regressors. In *International Conference on Machine Learning*, pages 4702–4712. PMLR, 2020.

Akira Kagawa and Steven W Leavitt. Stable carbon isotopes of tree rings as a tool to pinpoint the geographic origin of timber. *Journal of Wood Science*, 56(3):175–183, 2010.

Kaushi ST Kanankege, Moh A Alkhamis, Nicholas BD Phelps, and Andres M Perez. A probability co-kriging model to account for reporting bias and recognize areas at high risk for zebra mussels and eurasian watermilfoil invasions in minnesota. *Frontiers in veterinary science*, 4:231, 2018.

Jason Kirk, Joaquin Ruiz, John Chesley, Spencer Titley, and Spence Titley. The origin of gold in south africa: Ancient rivers filled with gold, a spectacular upwelling of magma and a colossal meteor impact combined to make the witwatersrand basin a very special place. *American Scientist*, 91(6):534–541, 2003.

Geoff Koehler, Kevin J Kardynal, and Keith A Hobson. Geographical assignment of polar bears using multi-element isoscapes. *Scientific Reports*, 9(1):1–9, 2019.

Andreas Krause and Carlos Guestrin. Nonmyopic active learning of gaussian processes: an

exploration-exploitation approach. In *Proceedings of the 24th international conference on Machine learning*, pages 449–456, 2007.

Kailin Kroetz, Gloria M Luque, Jessica A Gephart, Sunny L Jardine, Patrick Lee, Katrina Chicojay Moore, Cassandra Cole, Andrew Steinkruger, and C Josh Donlan. Consequences of seafood mislabeling for marine populations and fisheries management. *Proceedings of the National Academy of Sciences*, 117(48):30318–30323, 2020.

Naoki Kurashima, Yukiko Makino, Setsuko Sekita, Yasuteru Urano, and Tetsuo Nagano. Determination of origin of ephedrine used as precursor for illicit methamphetamine by carbon and nitrogen stable isotope ratio analysis. *Analytical chemistry*, 76(14):4233–4236, 2004.

Jin Li and Andrew D Heap. A review of spatial interpolation methods for environmental scientists. 2008.

Chao Ma, Hannah B Vander Zanden, Michael B Wunder, and Gabriel J Bowen. assignr: an r package for isotope-based geographic assignment. *Methods in Ecology and Evolution*, 11(8): 996–1001, 2020.

J.M. Muñoz-Redondo, D. Bertoldi, A. Tonon, L. Ziller, F. Camin, and J.M. Moreno-Rojas. Tracing the geographical origin of spanish mango (mangifera indica l.) using stable isotopes ratios and multi-element profiles. *Food Control*, 125:107961, 2021. ISSN 0956-7135. doi: https://doi.org/10.1016/j.foodcont.2021.107961. URL https://www.sciencedirect.com/science/article/pii/S0956713521000992.

NASA Earth Observations / NEO. Reflected shortwave radiation data, 2023. URL https://neo.gsfc.nasa.gov/view.php?datasetId=CERES_SWFLUX_M.

Sarah Pederzani and Kate Britton. Oxygen isotopes in bioarchaeology: Principles and applications, challenges and opportunities. *Earth-Science Reviews*, 188:77–107, 1 2019. ISSN 00128252. doi: 10.1016/j.earscirev.2018.11.005.

Naren Ramakrishnan, Chris Bailey-Kellogg, Satish Tadepalli, and Varun N Pandey. Gaussian processes for active data mining of spatial aggregates. In *Proceedings of the 2005 SIAM International Conference on Data Mining*, pages 427–438. SIAM, 2005.

Saeida Saadat, Hardi Pandya, Aayush Dey, and Deepak Rawtani. Food forensics: Techniques for authenticity determination of food products. *Forensic Science International*, page 111243, 2022.

Nele Schmitz, Volker Haag, Céline Blanc-Jolivet, Markus Boner, María Teresa Cervera, Manuel Chavesta, Richard Cronn, Victor Deklerck, Carmen Diaz-Sala, Eleanor Dormontt, Peter Gasson, David Gehl, John C. Hermanson, Eurídice Honorio Coronado, Cady Lancaster, Frederic Lens, Estephanie Patricia Liendo Hoyos, Sandra Martínez-Jarquín, Rolando Antonio Montenegro, Kathelyn Paredes Villanueva, Tereza Cristina Monteiro Pastore, Tahiana Ramananantoandro, Harisoa Ravaomanalina, Alexandre Magno Sebbenn, Niklas Tysklind, Mart Vlam, Charlie Watkinson, and Michael Wiemann. General sampling guide for timber tracking. How to collect reference samples for timber identification. *Global Timber Tracking Network*, 2019.

Rolf TW Siegwolf, J Renée Brooks, John Roden, and Matthias Saurer. *Stable isotopes in tree rings: inferring physiological, climatic and environmental responses*. Springer Nature, 2022.

Anthony J. Silva, Rosalee S. Hellberg, and Robert H. Hanner. Chapter 7 - seafood fraud. In

Rosalee S. Hellberg, Karen Everstine, and Steven A. Sklare, editors, *Food Fraud*, pages 109–137. Elsevier, 2021. ISBN 978-0-12-817242-1.

Katie St. John Glew, Laura J Graham, Rona AR McGill, and Clive N Trueman. Spatial models of carbon, nitrogen and sulphur stable isotope distributions (isoscapes) across a shelf sea: An inla approach. *Methods in Ecology and Evolution*, 10(4):518–531, 2019.

Chang Su and Sargur Srihari. Evaluation of rarity of fingerprints in forensics. *Advances in Neural Information Processing Systems*, 23, 2010.

H Swofford and C Champod. Probabilistic reporting and algorithms in forensic science: stakeholder perspectives within the american criminal justice system. *Forensic Science International: Synergy*, 4:100220, 2022.

Nikolaas J Van der Merwe, JA Lee-Thorp, JF Thackeray, A Hall-Martin, FJ Kruger, H Coetzee, RHV Bell, and M Lindeque. Source-area determination of elephant ivory by isotopic analysis. *Nature*, 346(6286):744–746, 1990.

Peter van der Sleen, Pieter A Zuidema, and Thijs L Pons. Stable isotopes in tropical tree rings: theory, methods and applications. *Functional Ecology*, 31(9):1674–1689, 2017.

Hannah B. Vander Zanden, Anton D. Tucker, Kristen M. Hart, Margaret M. Lamont, Ikuko Fujisaki, David S. Addison, Katherine L. Mansfield, Katrina F. Phillips, Michael B. Wunder, Gabriel J. Bowen, Mariela Pajuelo, Alan B. Bolten, and Karen A. Bjorndal. Determining origin in a migratory marine vertebrate: a novel method to integrate stable isotopes and satellite tracking. *Ecological Applications*, 25(2):320–335, 2015.

Charles J Watkinson, Peter Gasson, Gareth O Rees, and Markus Boner. The development and

use of isoscapes to determine the geographical origin of quercus spp. in the united states. *Forests*, 11(8):862, 2020.

Charles J Watkinson, Gareth O Rees, Moundounga Cynel Gwenael, Peter Gasson, Sabine Hofem, Lina Michely, and Markus Boner. Stable isotope ratio analysis for the comparison of timber from two forest concessions in gabon. *Frontiers in Forests and Global Change*, page 155, 2022a.

Charles J Watkinson, Gareth O Rees, Sabine Hofem, Lina Michely, Peter Gasson, and Markus Boner. A case study to establish a basis for evaluating geographic origin claims of timber from the solomon islands using stable isotope ratio analysis. *Frontiers in Forests and Global Change*, 4, 2022b. ISSN 2624-893X.

Jason B West, Gabriel J Bowen, Todd E Dawson, and Kevin P Tu. *Isoscapes: understanding movement, pattern, and process on Earth through isotope mapping*. Springer, 2010.

Christopher KI Williams and Carl Edward Rasmussen. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA, 2006.

Andrew Gordon Wilson, Zhiting Hu, Ruslan Salakhutdinov, and Eric P Xing. Deep kernel learning. In *Artificial intelligence and statistics*, pages 370–378. PMLR, 2016.

Barry T Wilson, Andrew J Lister, Rachel I Riemann, and Douglas M Griffith. Live tree species basal area of the contiguous united states (2000-2009). 2013.

Michael B Wunder. Using isoscapes to model probability surfaces for determining geographic origins. *Isoscapes: understanding movement, pattern, and process on Earth through isotope mapping*, pages 251–270, 2010.

Table 1: Mean test set performance for all the models used in the study. Best values across all models are shown in bold and underlined whereas values that are not significantly different from the best values (Wilcoxon signed-rank test, p=0.05) are shown in bold. The Spatial-only GP combined with the density prior gives the highest predictive log-likelihood and log-posterior and the lowest MAE and ASH values for all priors used. The Spatial-only model outperforms the other models when range or density priors are used, while the Atmospheric+Spatial model performs best in terms of MAE and ASH when flat priors are used. The inclusion of species distribution information decreases MAE and ASH values for all models used. All of our models outperform the earlier method of Watkinson *et al.* (Watkinson et al., 2020) on most or all metrics.

model	prior	log L	mode distance (km)	MAE (km)	log-posterior	ASH (km^2)
Spatial-only	flat	**-6.964**	**433**	809	-9.582	470000
Spatial-only	range	**-6.964**	**435**	600	-9.537	327000
Spatial-only	density	**-6.964**	**400**	**520**	**-9.059**	**203000**
Atmospheric-only	flat	-7.362	531	870	-9.972	576000
Atmospheric-only	range	-7.362	505	606	-9.797	450000
Atmospheric-only	density	-7.362	534	567	-9.428	311000
Atmospheric+Spatial	flat	**-7.149**	**408**	794	-9.518	382000
Atmospheric+Spatial	range	**-7.149**	**399**	627	-9.431	315000
Atmospheric+Spatial	density	**-7.149**	**463**	536	**-8.978**	**213000**
Watkinson *et al.*	NA	NA	886	859	NA	691000

Michael B Wunder and Ryan D Norris. Improved estimates of certainty in stable-isotope-based methods for tracking migratory animals. *Ecological applications*, 18(2):549–559, 2008.

Stefan Ziegler, Stefan Merker, Bruno Streit, Markus Boner, and Dorrit E Jacob. Towards understanding isotope variability in elephant ivory to establish isotopic profiling and source-area determination. *Biological Conservation*, 197:154–163, 2016.

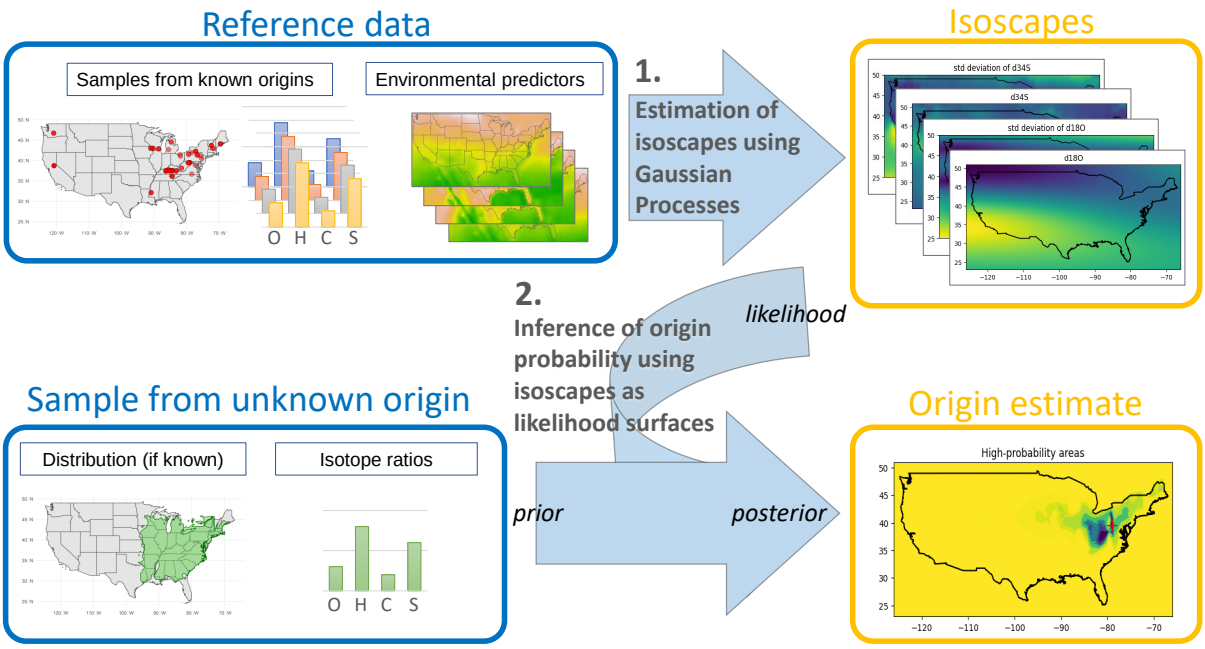
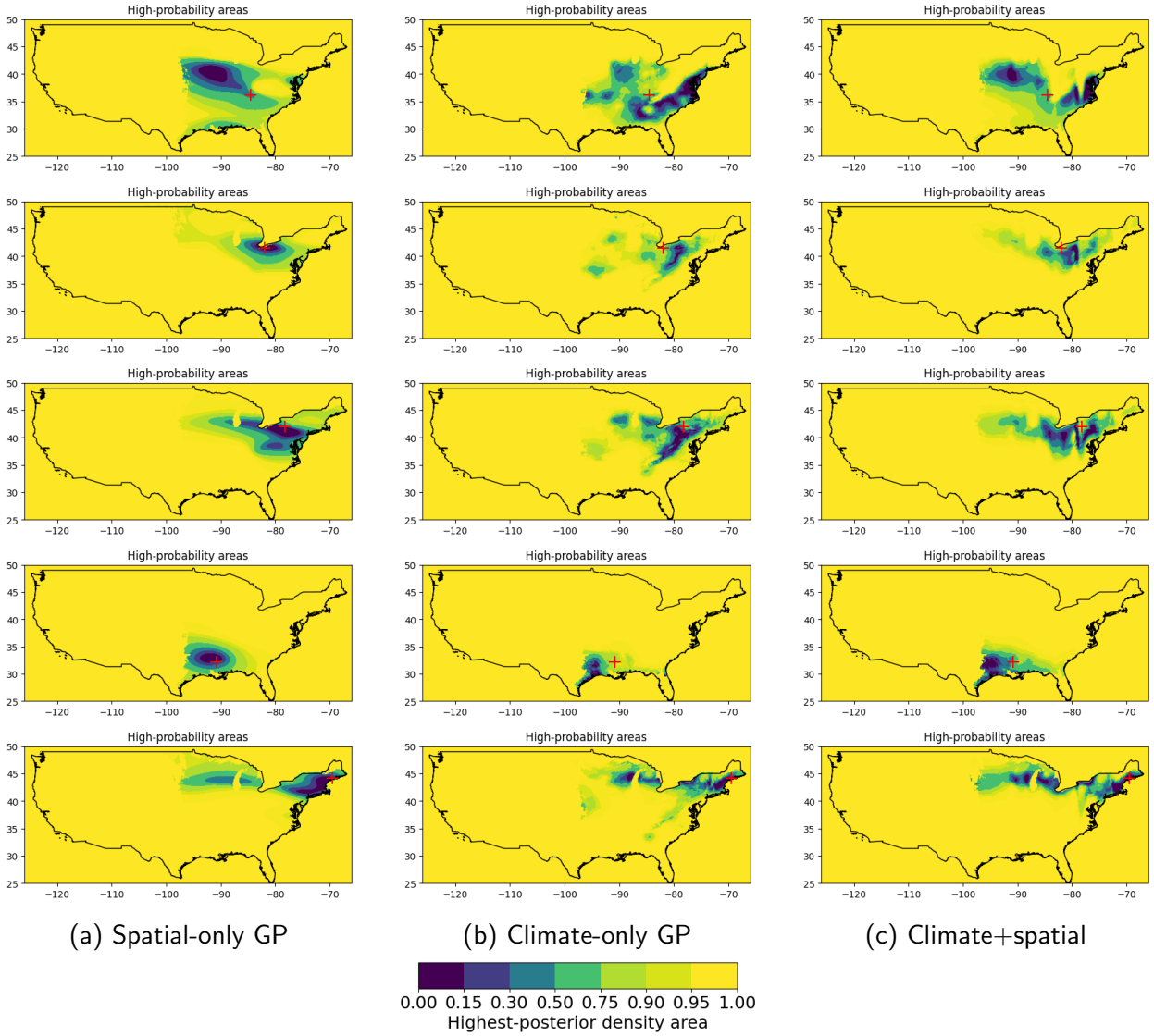
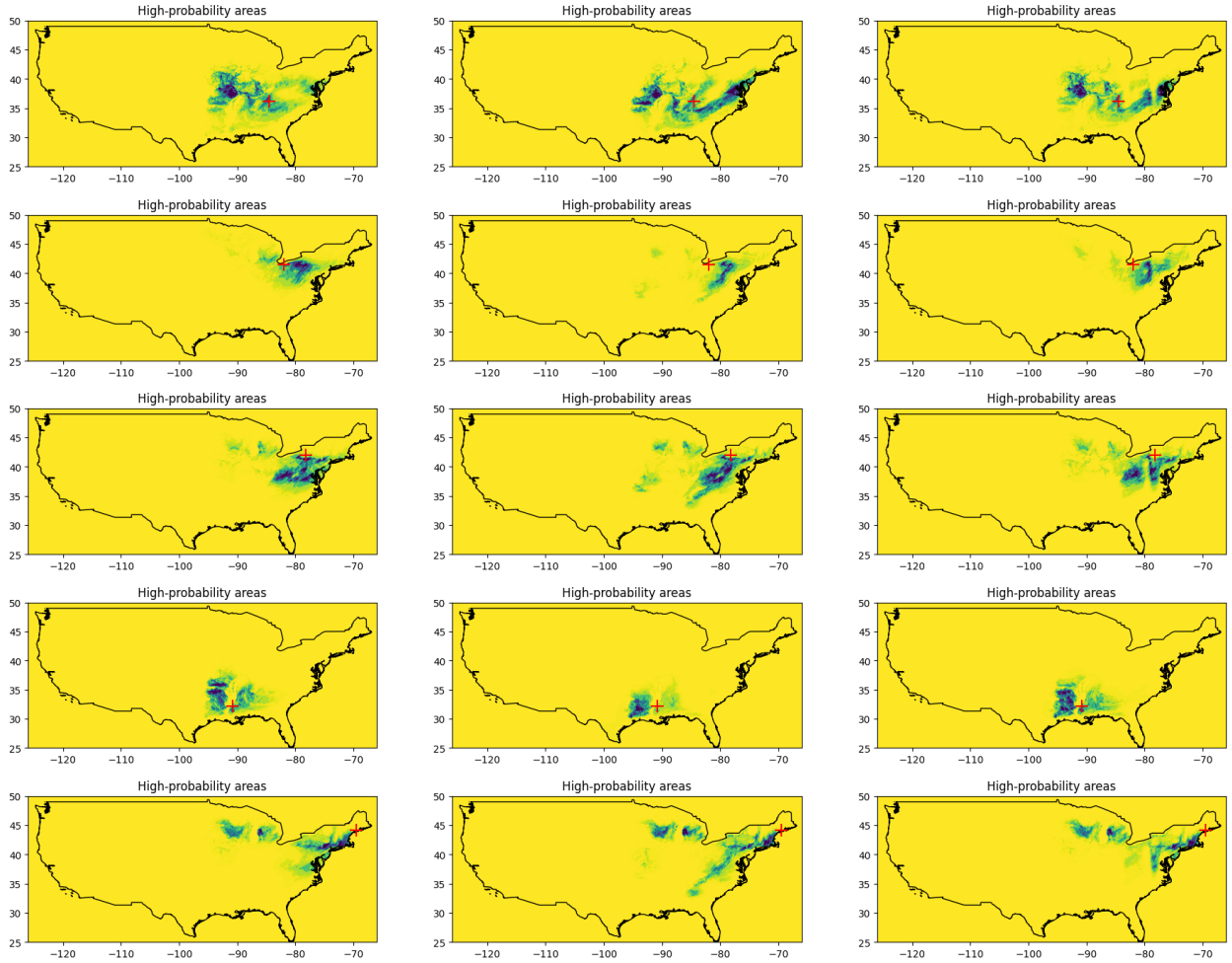


Figure 1: Model workflow. We use a training set of isotope ratios from trees collected at known locations and atmospheric data layers ("Reference data"). We fit a Gaussian process regression model to infer isoscapes and associated variance estimates, and compute the likelihood of observing the IR value for each element across the study area. To estimate the source of material with uncertain provenance ("Samples from unknown origin"), the isoscapes are then combined with prior information on the geographical distribution of the species, to yield a probability distribution of origin for the sample. We visualize predicted probability maps by plotting highest-posterior density regions for several probability levels (15%, 30%, 50%,75%,90% and 95%, dark blue to light green).



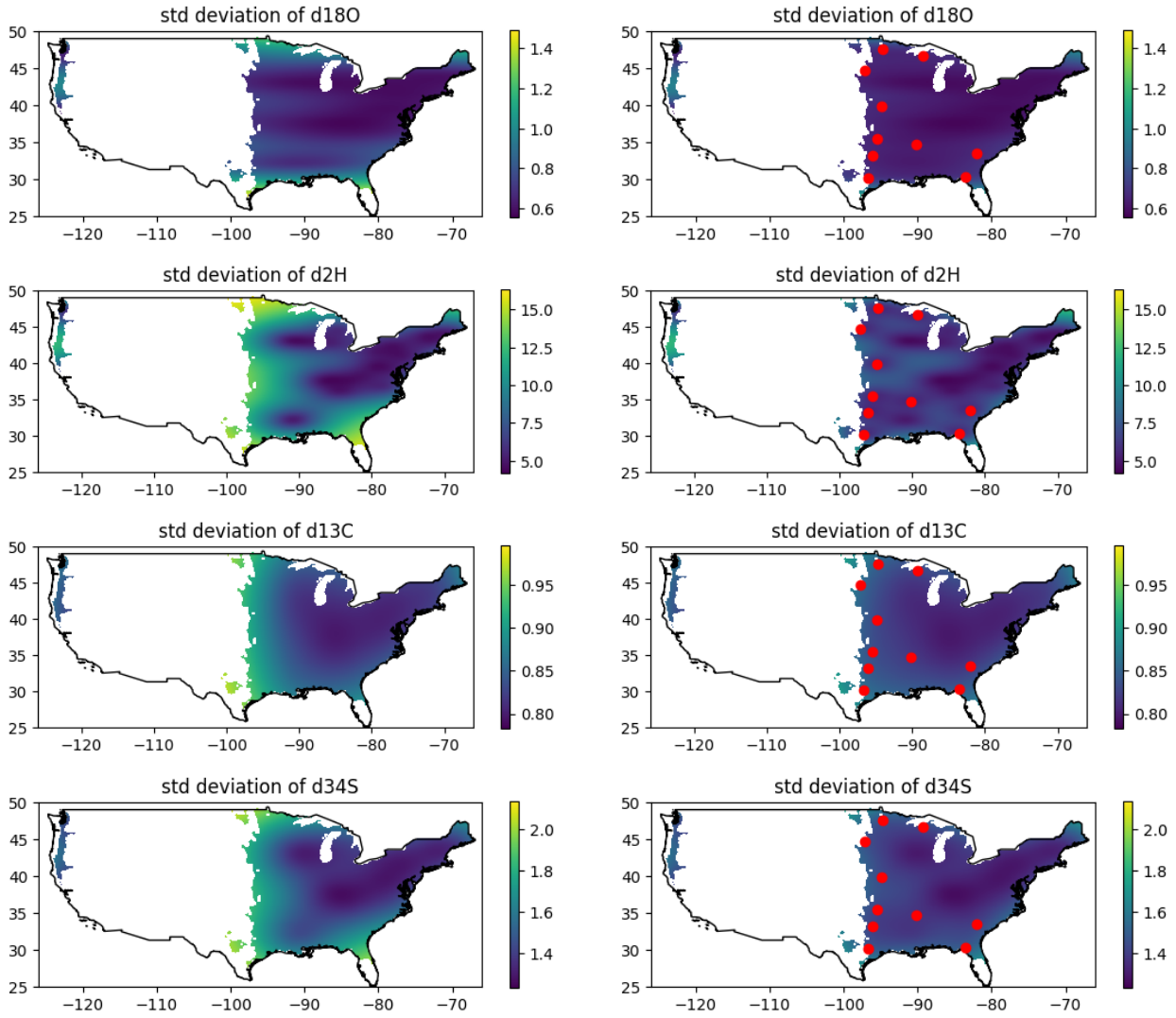
(a) Spatial-only GP (b) Climate-only GP (c) Climate+spatial

Figure 2: Harvest location predictions from the three models for 5 points from the test set using the range prior. Darker shades denote areas with higher posterior probability with thresholds set so that the total probability of the colored area is equal to a specified value (see color chart). The red cross indicates the actual location of the tree.



(a) Spatial-only GP (b) Climate-only GP (c) Climate+spatial

Figure 3: Harvest location predictions from the three models for 5 points from the test set using the density prior. Darker shades denote areas with higher posterior probability with thresholds set so that the total probability of the colored area is equal to a specified value (see color chart). The red cross indicates the actual location of the tree.



(a) Before active learning

(b) After active learning

Figure 4: Maps showing predictive standard deviations for the four isotopes before and after adding 10 sample locations proposed by our active learning method for the spatial-only model. Standard deviations are only shown within the allowed sampling area, which is the union of ranges for the species in our data set. The red dots show the proposed locations. Our method proposes locations in areas with high predictive variance, particularly for δ^2H and δ^{34}S. Adding the proposed locations leads to a marked reduction of variance in the neighboring areas.

A More detail on Gaussian Processes

In the following, we give a brief overview of GPs. For an in-depth discussion, see Williams and Rasmussen (Williams and Rasmussen, 2006).

GPs provide a flexible framework for regression, which enables the modeler to quantify the uncertainty of specific inferences. A GP is a random process such that all of its marginals are jointly normally distributed (Gaussian). Let $\mathbf{x} = [x_{lon}, x_{lat}]$ be the GPS coordinates of a sample. For any set of positions $\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_n}$, the responses $y_1, y_2 \ldots, y_n$ at those positions are assumed to be jointly normally distributed

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} m(\mathbf{x_1}) \\ m(\mathbf{x_2}) \\ \vdots \\ m(\mathbf{x_n}) \end{bmatrix}, \begin{bmatrix} k(\mathbf{x_1}, \mathbf{x_1}) & k(\mathbf{x_1}, \mathbf{x_2}) & \ldots & k(\mathbf{x_n}, \mathbf{x_n}) \\ k(\mathbf{x_2}, \mathbf{x_1}) & k(\mathbf{x_2}, \mathbf{x_2}) & \ldots & k(\mathbf{x_2}, \mathbf{x_n}) \\ \vdots & \vdots & \ddots & \vdots \\ k(\mathbf{x_n}, \mathbf{x_1}) & k(\mathbf{x_n}, \mathbf{x_2}) & \ldots & k(\mathbf{x_n}, \mathbf{x_n}) \end{bmatrix} + \sigma^2 \mathbf{I} \right) \tag{A.2}$$

where:

1. The *mean function* $m(\mathbf{x})$ describes the *a priori* expected value of the response y at location $\mathbf{x}$. We use the constant mean $m_c(\mathbf{x}) = \theta_m$ for all $\mathbf{x}$, where θ_m is a parameter to be estimated from the data.

2. The *covariance function* $k(\mathbf{x_1}, \mathbf{x_2})$ describes the *a priori* covariance between responses at locations $\mathbf{x_1}$ and $\mathbf{x_2}$. This is also a parameterized function. Popular choices of k are the *squared exponential* $k_{se}(\mathbf{x_1}, \mathbf{x_2}) = A \exp(-|\mathbf{x_1} - \mathbf{x_2}|^2/\rho^2)$, or the Matern function (Williams and Rasmussen, 2006), which both reflect the common assumption that similar predictor values will lead to similar response values. In this work, we use the Matern function with separate scaling parameters for latitude and longitude to model

spatial covariance.

3. The noise parameter σ^2 models measurement error.

4. $\mathbf{I}$ is the $n \times n$ identity matrix.

We write $\mathbf{y}$, $\mathbf{m}$ and K to denote the responses, means and the covariance matrix of the training data, respectively, so that we can write Eq. A.2 as $\mathbf{y} \sim \mathcal{N}(\mathbf{m}, K + \sigma^2\mathbf{I})$. The choice of mean and covariance functions reflects prior knowledge and modelling assumptions about the regression problem. The function parameters as well as the noise parameter σ are estimated by maximizing the likelihood of the training data. We use GPyTorch (Gardner et al., 2018) to efficiently find the maximum likelihood parameter estimates.

After parameter estimation, the GP regression model can be used to predict responses at previously unseen data points. Let S be the locations and responses comprising the training data set. Since the responses at training and test points are assumed to be jointly Gaussian, the conditional distribution of the response at a test point $\mathbf{x}^*$ given the training data is also Gaussian with mean

$$\mu(\mathbf{x}^*|S) = m(\mathbf{x}^*) + \mathbf{k}^*(K + \sigma^2\mathbf{I})^{-1}(\mathbf{y} - \mathbf{m}) \tag{A.3}$$

where $\mathbf{k}^* = [k(\mathbf{x}^*, \mathbf{x}_1), k(\mathbf{x}^*, \mathbf{x}_2), \ldots, k(\mathbf{x}^*, \mathbf{x}_n)]$ is the vector of *a priori* covariances between responses at $\mathbf{x}^*$ and training data points. The posterior variance of y^* is given by

$$\sigma^2(\mathbf{x}^*|S) = k(\mathbf{x}^*, \mathbf{x}^*) + \sigma^2 - \mathbf{k}^*(K + \sigma^2\mathbf{I})^{-1}\mathbf{k}^{*\top} \tag{A.4}$$

- see (Williams and Rasmussen, 2006) for a derivation.

For a specific response value y^+, its likelihood of being observed at $\mathbf{x}^*$ is just the Gaussian

probability density with mean μ and variance σ^2 found by applying Equations A.3 and A.4

$$p(y^* = y^+|\mathbf{x}^*, S) = \frac{1}{\sqrt{2\pi\sigma^2(\mathbf{x}^*|S)}} \exp\left(\frac{-(y^+ - \mu(\mathbf{x}^*|S))^2}{2\sigma^2(\mathbf{x}^*|S)}\right) \tag{A.5}$$

For a sample $\bar{y} = (y_O, y_H, y_C, y_S)$ of observed isotope ratio values (meaning δ^{18}O, δ^{2}H, δ^{13}C, δ^{34}S) , the Bayes' theorem gives the posterior distribution of possible harvest locations:

$$p(\mathbf{x}|\bar{y}, S) = \frac{p(\mathbf{x}) \prod_{i \in \{O,H,C,S\}} p_i(y_i|\mathbf{x}, S)}{\int_{\mathbf{x} \in A} p(\mathbf{x}) \prod_{i \in \{O,H,C,S\}} p_i(y_i|\mathbf{x}, S) d\mathbf{x}} \tag{A.6}$$

where the probabilities p_i are computed from the GP models for the respective isotopes using Equation A.5 and A is the study area. The integral in the denominator is computed by averaging the probabilities over the spatial grid.

B Active learning performance

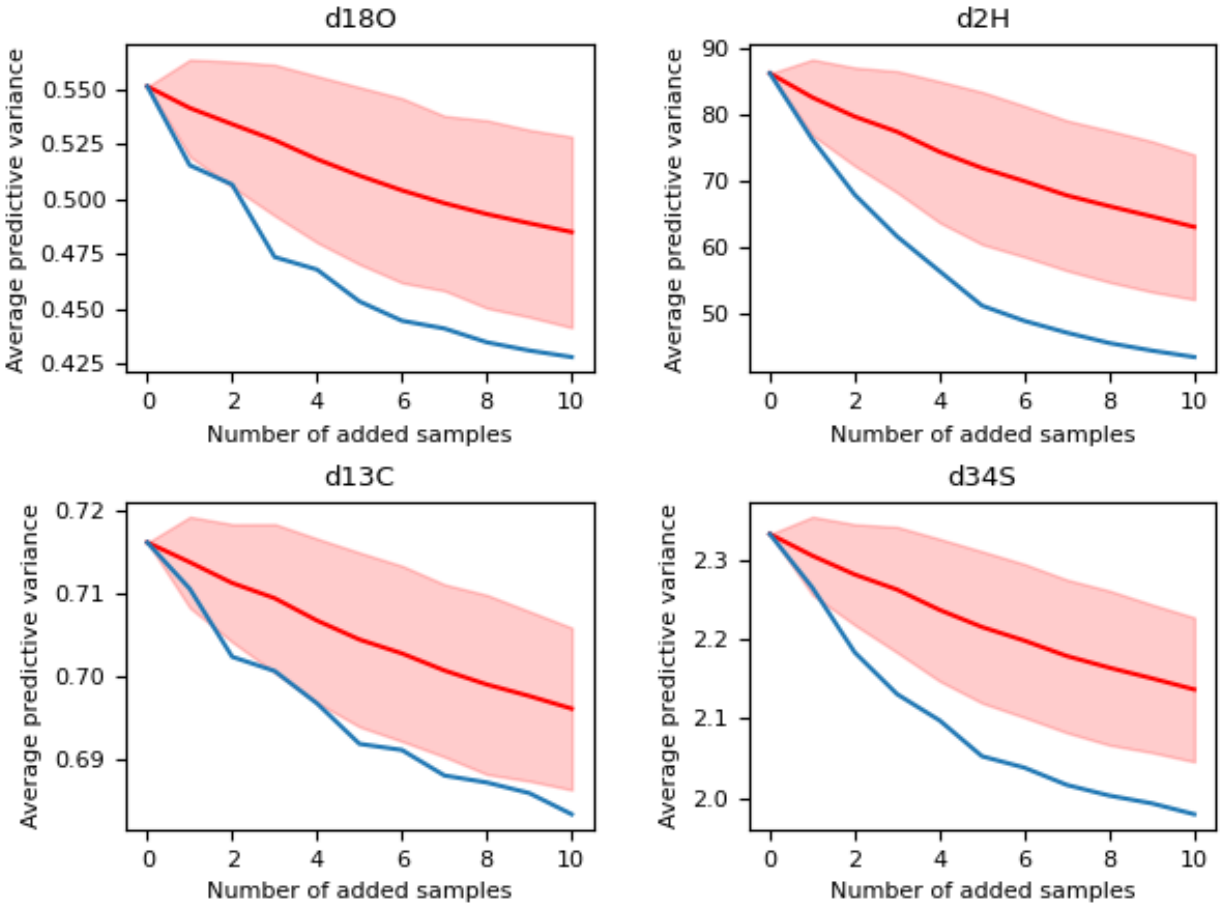


Figure 5: Average predictive variances for $\delta^{18}O$,$\delta^{2}H$,$\delta^{13}C$ and $\delta^{34}S$ as a function of the number of samples added to the base training data set; blue - active learning strategy; red - random sampling (shaded area denotes values within two standard deviations of the mean across $n_r = 100$ simulations).