A machine learning approach to estimating the geographical origin of timber

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, Boulder, Colorado, 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

February 27, 2023

Abstract

1: Determining the harvest location of timber is crucial to enforcing international regulations designed to tackle illegal logging and associated trade in forest products. However, complex supply chains obscure harvest sources, which often leaves paper-based traceability systems as the sole tool for demonstrating provenance, despite its vulnerability to fraud. 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 SIR predicted from reference data ('isoscapes'). The primary challenge in developing reliable isoscapes is the need to accurately predict stable isotopes in areas where no physical reference samples are available. Existing attempts to predict isoscapes from reference data have been hampered by the use of simple and ad-hoc statistical models, limiting the precision of estimated isoscapes and the confidence in derived estimates of geographical origin.

2: We present a new SIRA data analysis pipeline, designed to infer timber harvest location. We use Gaussian Processes to robustly estimate isoscapes from reference wood samples, which are then combined with species distribution range data to compute, for every pixel in the study area, the probability of it being the origin of the sample. Finally we present a methodology to determine priority locations to obtain new reference samples in future field expeditions.

3: We demonstrate our approach on a data set of $n = 87$ wood samples from seven oak species in the USA as proof of concept. Our method is able to determine the harvest location up to 520-870 km, depending on the model parameterisation. Incorporating species distribution information improves accuracy by up to 36%. The new sampling locations proposed by our method decrease the variance of resultant isoscapes by up to 86% more than sampling the same number of locations at random.

4: The method we present here combines the prediction of isoscapes with derivation of geographical origin estimates. It advances the toolset available to authorities addressing illegal trade in forest products and enforcing anti-deforestation legislation. Importantly, reference data can be added as available, allowing for the expansion of reference collections and increasing prediction accuracy.

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

1 Introduction

One million species now face extinction, and the unsustainable exploitation of natural resources is the second largest driver of terrestrial biodiversity loss, next only to land use changes [19]. To prevent our societies triggering a new wave of mass extinctions [3], nearly 200 nations have recently agreed on a new set of targets and goals under the Kunming-Montreal Global Biodiversity Framework. Under this international agreement, human-driven species extinctions must halt by 2030, in order to allow an appropriate level of natural recovery by 2050. In particular, Target 5 of the agreement has the objective to "ensure that the use, harvesting and trade of wild species is sustainable, safe and legal, preventing overexploitation".

Meeting such an ambitious but necessary target will require overcoming a key element of unsustainable use of natural resources: the illegal harvest of threatened trees. To combat illegal logging and associated trade in illegally harvested timber, various frameworks have been put into place, such as the Convention on the International Trade in Endangered Species (CITES) and security linked sanctions. At national or regional levels, additional legislation include the US Lacey Act, the UK Timber Regulation, the EU Deforestation Regulation and the Australian Illegal Logging Prohibition Act. Despite the comprehensive legal framework already in place, and the international commitments under current adoption, it is clear that none of these can effectively work without enforcement. This is where technological solutions and methodological developments become key.

1.1 Stable Isotope Ratio Analysis for timber provenance

Current legislation increasingly requires accurate, cost-effective, and high-throughput tools that can verify the specific species and origin of products in global trade [20]. Scientific testing technologies have become relatively well established and are already supporting companies and enforcement authorities to scrutinize traceability systems. These technologies measure the chemical, anatomical and genetic features of plants which, when compared against a robust physical reference collection, can be used to (in-)validate declared species and origin claims and support enforcement officials in their efforts to detect, for example, illegal or conflict timber and fraud in supply chains.

One of the most promising and widely used scientific technologies is stable isotope ratio analysis (SIRA). The ratios of several elemental stable isotopes within natural products vary across space and can assist in verifying the geographic origin of products. Most of the abundant elements in organic compounds (Hydrogen, Oxygen, Carbon, Sulfur, Nitrogen) have naturally-occurring stable isotopes that do not undergo radioactive decay, and can be readily detected by mass spectrometry [5]. The composition of stable isotopes incorporated into the tissues of a plant is determined by the soil, climate, metabolic fractionation and other biotic and abiotic conditions characteristic of the species and its habitat [41, 9, 26]. Stable isotope ratios can be used to discriminate between geographic areas as the variation in these stable isotope ratios depend on natural variations of the underlying mechanisms (for example environmental drivers [47]). The types of products that can be analysed by SIRA to determine risk for being illegally harvested include natural resources-such as forest products [48, 5], agricultural products [10, 39], wildlife [8, 32], fish/seafood [16, 42, 34], ivory [46, 54], precious metals [31] and illicit drug trafficking, including natural and synthetic opioids [35, 11], and other forensic uses to identify counterfeit and pirated goods trafficking [12].

1.2 Modelling approach

Current modelling practices for the use of SIRA to verify harvest location of both legally and illegally harvested products could be improved. The use of SIRA is currently limited by the relative simplicity of models used, as well as by the limited number of reference samples used as input data for such models. The practical nature of reference sampling campaigns is that they can be 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 [40]. There has been considerable development of isoscapes ("isotope landscapes"), given that stable isotopic variation is a continuous spatial variable in nature [50]. These isoscapes are geospatial maps that show the isotopic variation of the material of interest [50]. While the potential of isoscapes for determining 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 [48, 49] and clustering [32], which do not fully leverage the information contained in isotope ratio data.

Gaussian Process (GP) regression, also known as kriging in geostatistics literature, is a class of flexible regression models which use the values in sampled points to estimate the values in surrounding, unsampled points [36, 18, 51]. 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. Quantifying the uncertainty of predictions is viewed as increasingly important in safety-critical [28] and forensic [44, 45] machine learning applications. GP regression also facilitates co-kriging: inferring the values of a sparsely sampled variable of interest through variables that are highly correlated with it but more densely sampled [2, 30]. In the context of plant origin estimation, co-kriging 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 [26, 41]. This then provides a powerful tool for predicting the isotopic composition in areas that have not yet been sampled (examples are given in [23, 48]).

Previous work on isoscapes used GP regression primarily as a spatial interpolation technique without a probabilistic interpretation [23, 32, 30]. A more recent method uses GP variance estimates from precipitation isoscapes for origin determination in animals [37]. In this work, we develop GP-based probabilistic machine learning models to infer the origin of timber samples by directly modelling timber isoscapes. We present a new data analysis pipeline that incorporates timber SIRA data, atmospheric predictors and species distribution data. We find that probabilistic modeling greatly enhances the utility of SIRA in estimating the geographical origin of timber and helps guide future sample collection by identifying sampling locations that will minimize prediction uncertainty. The presented framework can then be applied to trace back timber of endangered species, by assisting in determining where to collect samples and by using SIRA datasets being collected by the World Forest ID [22] initiative.

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) [48]. Stable isotope ratio measurements were done following the protocol described in Boner et al. (2007) [5] and Watkinson et al. (2020) [48]. 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. Stable isotope ratios are largely driven by environmental conditions such as precipitation, temperature, humidity and so on. Thus, it is natural to use publicly available data on those factors to aid the inference of isoscapes. We used the following atmospheric data: δ^{2}H and δ^{18}O isotopic composition of precipitation [7], water vapour [6] (found to be associated with δ^{13}C by Watkinson *et al.* [48]), reflected shortwave radiation [1] and precipitation (multi-satellite) [27], both of which were found to be associated with δ^{34}S by Watkinson *et al.* [48].

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 [52], 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 $terra$ [25] to bilinearly aggregate abundance data so that it matched the spatial resolution of other spatial data in the pipeline.

2.2 Model architecture

Figure 5 presents the pipeline overview of the data sets, 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 Sections 2.3 and 2.4 for more details. The input to the GP consists of the coordinates and/or the climate variable values at the grid point. For a combination of stable isotope ratios (y_O, y_H, y_C, y_S) of observed stable isotope ratios, we compute the likelihood of observing it 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 perform Bayesian inference by computing the posterior probability of each grid point being the origin of the sample by applying Bayes' Theorem. 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%).

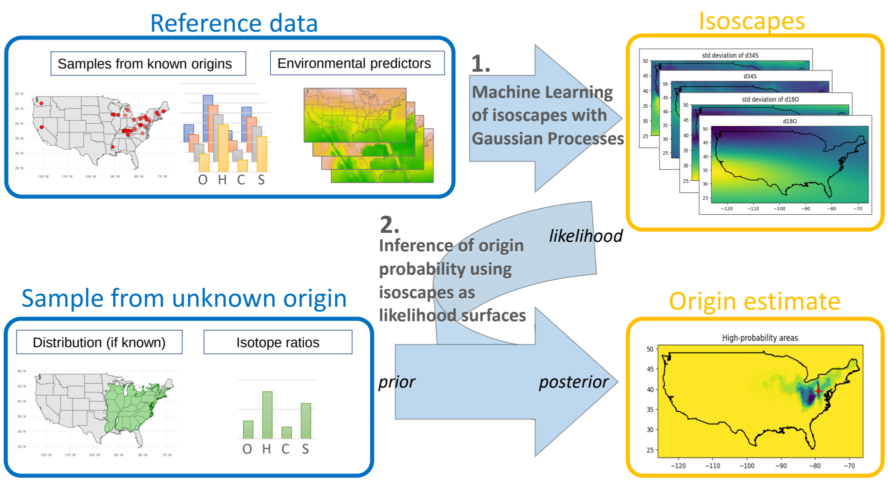


Figure 1: Model workflow. We use a training set of isotope ratios from trees collected at known locations and atmospheric data layers ("Reference data"). We fit a Gaussian process regression model to infer isoscapes and associated variance estimates, and compute the likelihood of observing the IR value 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).

2.3 Gaussian Process regression

In the following, we give a brief overview of GPs. For a more thorough explanation, see [51].

GPs provide a flexible framework for regression, which enables the modeller 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

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} m(\mathbf{x_1}) \\ m(\mathbf{x_2}) \\ \vdots \\ m(\mathbf{x_n}) \end{bmatrix}, \begin{bmatrix} k(\mathbf{x_1}, \mathbf{x_1}) & k(\mathbf{x_1}, \mathbf{x_2}) & \ldots & k(\mathbf{x_n}, \mathbf{x_n}) \\ k(\mathbf{x_2}, \mathbf{x_1}) & k(\mathbf{x_2}, \mathbf{x_2}) & \ldots & k(\mathbf{x_2}, \mathbf{x_n}) \\ \vdots & \vdots & \ddots & \\ k(\mathbf{x_n}, \mathbf{x_1}) & k(\mathbf{x_n}, \mathbf{x_2}) & \ldots & k(\mathbf{x_n}, \mathbf{x_n}) \end{bmatrix} + \sigma^2 \mathbf{I} \right) \quad (1)$$

where:

1. $m(\mathbf{x})$ is the *mean function* describing the *a priori* expected value of the response y at point $\mathbf{x}$. Usually, this is a parameterized function whose parameters are estimated by fitting the model to the training data. The most common choice of mean function is the constant mean $m_c(\mathbf{x}) = \theta_m$ for all $\mathbf{x}$, which we also use in this work.

2. $k(\mathbf{x_1}, \mathbf{x_2})$ is the *covariance function* describing the *a priori* covariance between responses at points $\mathbf{x_1}$ 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 [51], 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. σ^2 is the intrinsic noise parameter

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

We will 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. 1 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 [21] 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

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

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

$$\sigma^2(\mathbf{x}^*|S) = k(\mathbf{x}^*, \mathbf{x}^*) + \sigma^2 - \mathbf{k}^*(K + \sigma^2\mathbf{I})^{-1}\mathbf{k}^{*\top} \quad (3)$$

- see [51] for a derivation.

For a specific response value y^+, its likelihood at $\mathbf{x}^*$ is just the Gaussian probability density with mean $\mu$and variance σ^2 found by applying Equations 2 and 3

$$p(y^* = y^+|\mathbf{x}^*, S) = \frac{1}{\sqrt{2\pi\sigma^2(\mathbf{x}^*|S)}} \exp\left(\frac{-(y^+ - \mu(\mathbf{x}^*|S))^2}{2\sigma^2(\mathbf{x}^*|S)}\right) \quad (4)$$

2.4 Incorporating atmospheric data

For any $\mathbf{x}$, let $\mathbf{u}_i(\mathbf{x})$ denote the 12-entry vector of monthly values of atmospheric variable i at location $\mathbf{x}$. We use a linear covariance function to model the covariance component corresponding to the variation in atmospheric variable i

$$k_i(\mathbf{x_1}, \mathbf{x_2}) = \theta_i[\mathbf{u}_i(\mathbf{x_1})]^\top \mathbf{u}_i(\mathbf{x_2}) \quad (5)$$

where θ_i is a parameter to be estimated during training. The linear covariance function models a linear relationship between the atmospheric variable and the response and is mathematically equivalent to Bayesian linear regression [51].

The overall covariance function is the sum of the spatial term and the linear terms

$$k(\mathbf{x_1}, \mathbf{x_2}) = k_{spatial}(\mathbf{x_1}, \mathbf{x_2}) + \sum_{i \in V} \theta_i[\mathbf{u}_i(\mathbf{x_1})]^\top \mathbf{u}_i(\mathbf{x_2}) \quad (6)$$

where V is the set of atmospheric variables impacting the considered isotope ratio.

2.5 Bayesian inference of sample origin

Given a prior distribution $p(\mathbf{x})$ over possible sample origins and a GP regression model for every isotope, we perform Bayesian inference of sample origin. For a sample $\bar{y} = (y_O, y_H, y_C, y_S)$ of observed values, the Bayes' theorem gives the posterior distribution of possible origins:

$$p(\mathbf{x}|\bar{y}, S) = \frac{p(\mathbf{x}) \prod_{i \in \{O,H,C,S\}} p_i(y_i|\mathbf{x}, S)}{\int_{\mathbf{x} \in A} p(\mathbf{x}) \prod_{i \in \{O,H,C,S\}} p_i(y_i|\mathbf{x}, S) d\mathbf{x}} \quad (7)$$

where the probabilities p_i are computed from the GP models for the respective isotopes using Equation 4 and A is the study area. The integral in the denominator is computed by averaging the probabilities over the spatial grid.

2.6 Incorporating species distributions

We use the spatial density maps developed by Wilson *et al.* [52] 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.*.

2.7 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 n=74 test samples across the 5 folds.

Rigorously evaluating the performance of our model is a non-trivial task as it produces a distribution over possible locations, rather than a single location. For this reason, we have defined several metrics to investigate different aspects of our predictions.

2.7.1 Predictive log-likelihood and log-posterior

We report the log-likelihood (Eq. 4) and the log-posterior (Eq. 7) of observing the sample $\bar{y}$ at its true origin $\mathbf{x}_t$. Both of those measure how well the model fits the test data.

2.7.2 Mean posterior distance to true location (MPD)

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

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

where $d()$ is the great circle distance between the two points. A perfect prediction would have the distance of 0. In practice, isoscapes often predict similar isotope ratio values at distant locations, so even a statistically efficient method might yield a high MPD value. This metric will favour predictions concentrated around the true location over equally dispersed predictions concentrated elsewhere. It will also favour less dispersed predictions generally.

2.7.3 Area scored higher than the true location (ASH)

The behaviour of MPD 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.

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

where $I(.)$ is the indicator function that yields 1 when the statement is true and 0 otherwise. In contrast to MPD, 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.

2.8 Guiding future sampling efforts

Field sample collections are time-consuming and expensive. By having an idea which sample points need to be collected for increased origin estimation accuracy, we can optimize future field collections. 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 [15]. 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 [24, 38]. Here, we adopt an approach similar to that of Guestrin *et al.* [24] 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. Let S be the set of sampled locations and G be 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

$$IG(\mathbf{x}^*) = \sum_{\mathbf{x}\in G} \left[(\log(\sigma^2(\mathbf{x}|S)) - \log(\sigma^2(\mathbf{x}|S \cup \{\mathbf{x}^*\}))\right] \tag{8}$$

where the predictive variances are computed using Equation 3. 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 $\mathbf{x}^*$ in longitude or more than 7.5 degrees in latitude.

3 Results

3.1 Model accuracy and comparison

Table 1 shows performance metrics for all the models on the test data set. In all settings, the plausible origin areas identified by our models consisted of points within an average distance of 520-870 kilometers from the true locations. Even with a relatively small training data set of 69 training samples, our model is able to exclude the vast majority of the study area as a possible source of the sample under consideration.

Incorporating species distribution information improves prediction performance for every model and every metric examined except the log-likelihood, which does not depend on the prior. Informative priors improve MPD 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 predictions to the true location, except when a flat prior is used. In general, the spatial-only and the combined spatial+atmospheric model give similar results on all metrics and they both 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 data set size.

The predictions of atmospheric GP models appear qualitatively different from those from the purely spatial GP. Atmospheric model predictions often emphasize geographical areas with distinct climate patterns, such as the 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.

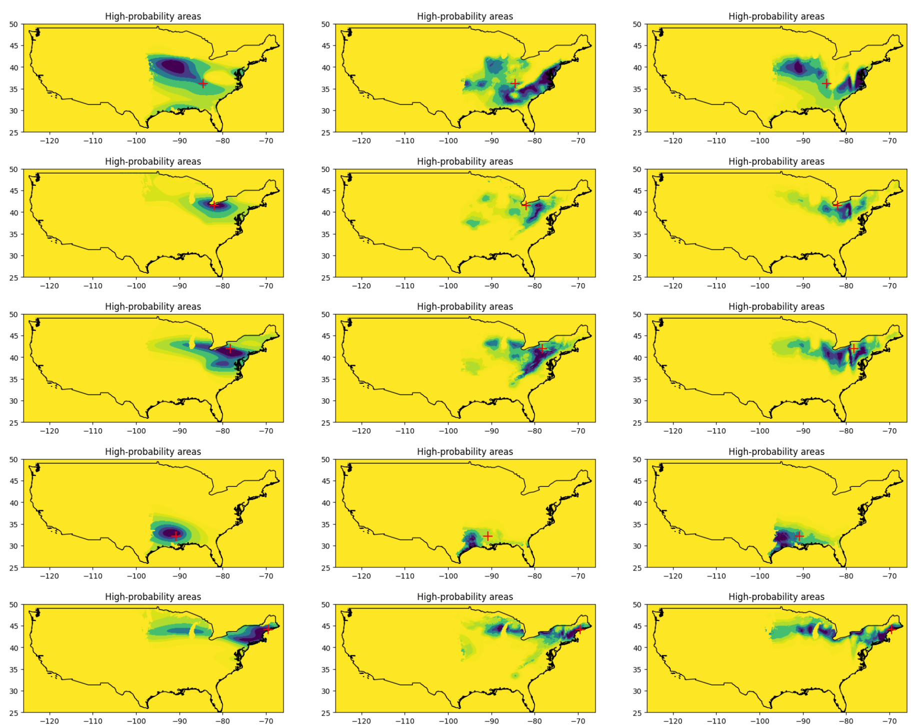
3.2 Guiding future sampling efforts

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 $n_s = 10$ 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

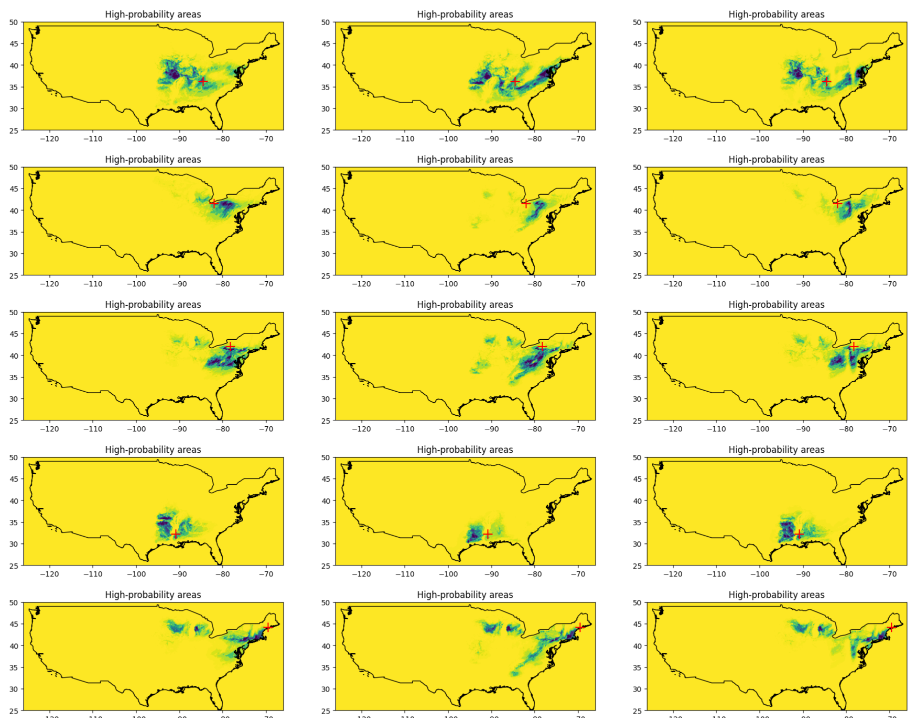
model	prior	log L	MPD (km)	log-posterior	ASH (km^2)
Spatial-only	flat	**-6.964**	809	-9.582	470000
Spatial-only	range	**-6.964**	600	-9.537	327000
Spatial-only	density	**-6.964**	**520**	-9.059	**203000**
Atmospheric-only	flat	-7.362	870	-9.972	576000
Atmospheric-only	range	-7.362	606	-9.797	450000
Atmospheric-only	density	-7.362	567	-9.428	311000
Atmospheric+Spatial	flat	-7.149	794	-9.518	382000
Atmospheric+Spatial	range	-7.149	627	-9.431	315000
Atmospheric+Spatial	density	-7.149	536	**-8.978**	213000

Table 1: Mean test set performance for all the models used in the study. Best values across all models 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 MPD 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 MPD and ASH when flat priors are used. The inclusion of species distribution information decreases MPD and ASH values for all models used.



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

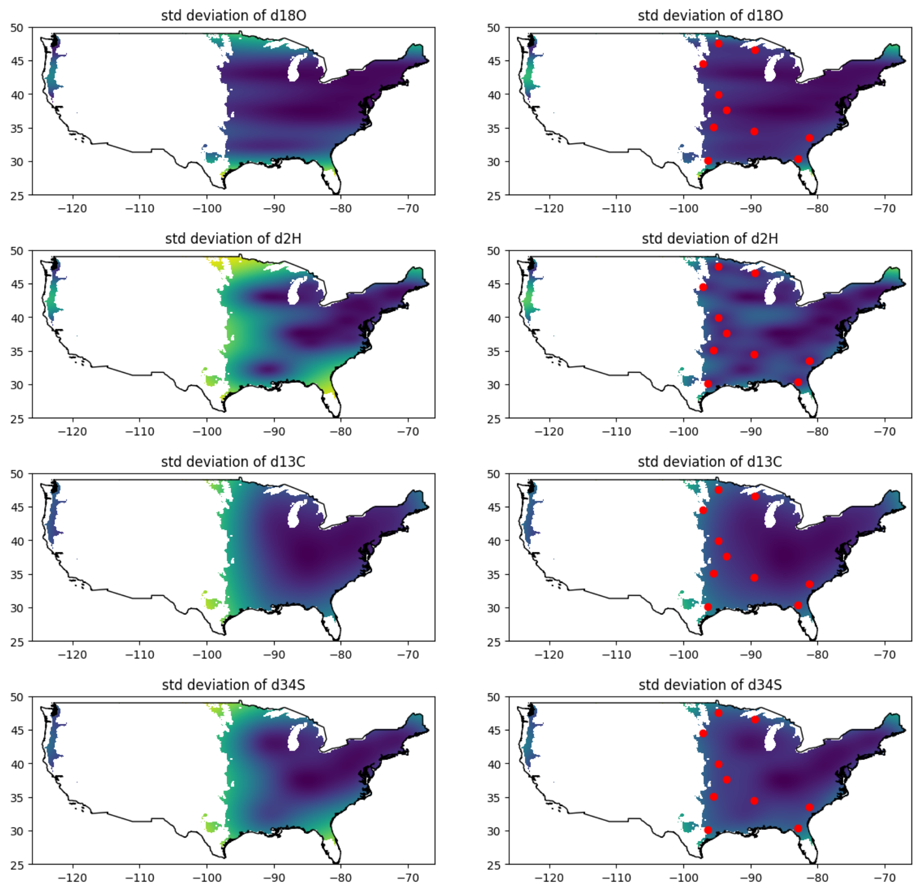
Figure 2: Spatial predictions from the three models for 5 points from the test set using the range prior. Darker shades denote areas with higher probability mass and the red cross indicates the actual location of the tree.



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

Figure 3: Spatial predictions from the three models for 5 points from the test set using the density prior. Darker shades denote areas with higher probability mass and the red cross indicates the actual location of the tree.

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.



(a) Before active learning (b) After active learning

Figure 4: Maps showing predictive standard deviations for the four isotopes before and after adding $n_s = 10$ 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.

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 $n_r = 100$ such variance maps and compared the average variance (across the allowed sampling area) of those maps with the maps in Fig. 4. Fig.S1 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 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 Timber origin estimation

To halt illegal logging, to enforce timber regulations and to protect biodiversity in forested landscapes, we need to be able to accurately estimate the timber harvest location. Although several examples exist of applying SIRA for timber origin questions [23, 48, 29], 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 this work we present a new computational pipeline which aims at taking advantage of both. The accuracy of our models depends on the specific modelling approach being used and the data sets incorporated. 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.* [48], 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. Like Watkinson *et al.* [48], our method 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.

Estimating the spatial covariance structure has recently attracted attention in animal stable isotope studies. Ma *et al.* [37] recently proposed a method that uses probabilistic precipitation isoscapes derived from a GP [14], which are then calibrated to produce isoscapes for the species of interest. St. John Glew *et al.* [43] 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 was first considered by Wunder [53].

Our current best performing model can estimate the origin of harvest location for *Quercus* species to 520 km across the (north-)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 mainly a focus on tropical species on which the logging pressure is significant and which might be endangered.

4.2 Guiding future collection efforts

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 (n=87) is quite small relative to the area of 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 [4], or deep GP models where the covariance function is modelled by a neural network [17].

Under the World Forest ID Programme [22], tens of thousands of tree samples are being collected globally, and are being analysed by different techniques, including SIRA. 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 [13], 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 [33], 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 cost of collecting samples as a function of the time 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, such as under the auspices of the World Forest ID Programme (https://worldforestid.org/), 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 Conflict of interest statement

The authors declare that they have no conflicts of interest.

7 Data availability

The data and code used in this study will be made publicly available within a year of publication. Earlier access may be provided by request.

8 Author contributions

Jakub Truszkowski, Victor Deklerck and Alexandre Antonelli jointly conceived the project. Jakub Truszkowski designed the methodology, implemented the algorithms, performed most of the analyses and wrote parts of the manuscript. Roi Maor selected and pre-processed some of the data sets, designed the main figure and wrote parts of the manuscript. Raquib Bin Yousuf, Subhodip Biswas, Naren Ramakrishnan and John Simeone wrote some of the software code and performed initial data explorations. Scot McQueen selected the data sets for species distribution models. Caspar Chater, Peter Gasson, Marigold Norman and Jade Saunders wrote parts of the manuscript. Victor Deklerck directed the project and wrote parts of the manuscript. All authors gave final approval for submission.

9 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 dataset. The findings and conclusions in the article are those of the authors.

References

[1] URL: `https://neo.gsfc.nasa.gov/view.php?datasetId=CERES_SWFLUX_M`.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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`.

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

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] Lesley A Chesson, Janet E Barnette, Gabriel J Bowen, Craig S Cook, Charles B Douthitt, John D Howa, Janet M Hurley, Helen W Kreuzer, Michael J Lott, Luiz A Martinelli, Shannon P O'Grady, David W Podlesak, Brett J Tripple, Luciano O Valenzuala, and Jason B West. Applying the principles of isotope analysis in plant and animal ecology to forensic science in the americas. *Oecologia*, 187(4):1077–1094, 2018.

[13] 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.

[14] 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.

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

[16] 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.

[17] Andreas Damianou and Neil D Lawrence. Deep gaussian processes. In *Artificial intelligence and statistics*, pages 207–215. PMLR, 2013.

[18] V. Deklerck. Timber origin verification using mass spectrometry: Challenges, opportunities, and way forward. *Forensic Science International: Animals and Environments*, 3:100057, 2023.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

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

[26] 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.

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

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

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

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

[37] 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.

[38] 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.

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

[40] 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.

[41] 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.

[42] 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.

[43] 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.

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

[45] 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.

[46] 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.

[47] 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.

[48] 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.

[49] 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, 2022.

[50] 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.

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

[52] 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.

[53] 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.

[54] 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.

Supplementary material

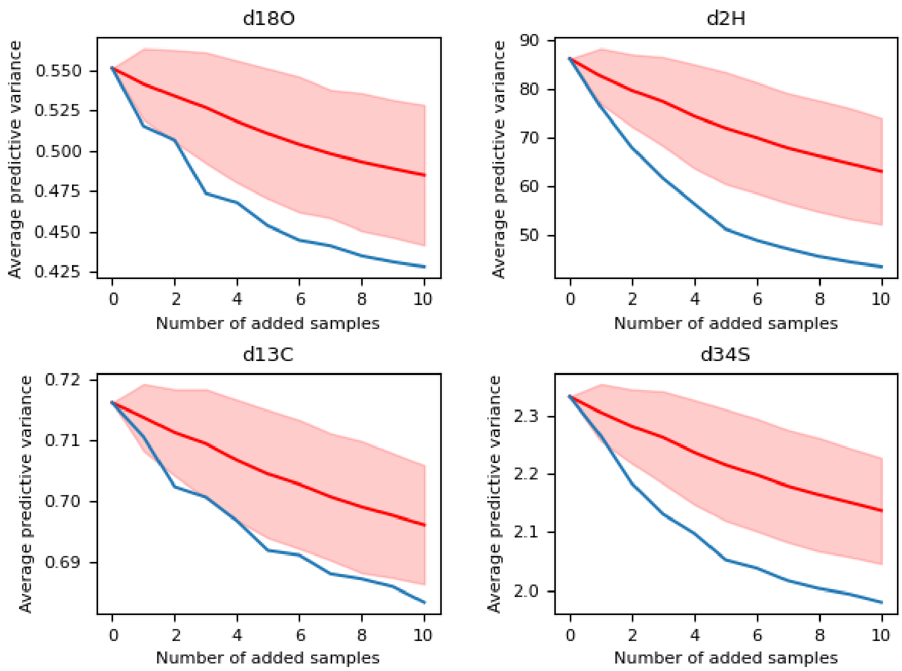


Figure 5: Average predictive variances for $\delta^{18}O$,$\delta^{2}H$,$\delta^{13}C$ and $\delta^{34}S$ as a function of the number of samples added to the base training data set; blue - active learning strategy; red - random sampling (shaded area denotes values within two standard deviations of the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