Chasing the Timber Trail: Machine Learning to Reveal Harvest Location Misrepresentation

Shailik Sarkar
Department of Computer Science
Virginia Tech
Alexandria, VA, USA
shailik@vt.edu

Raquib Bin Yousuf
Department of Computer Science
Virginia Tech
Alexandria, VA, USA
raquib@vt.edu

Linhan Wang
Department of Computer Science
Virginia Tech
Alexandria, VA, USA
linhan@vt.edu

Brian Mayer
Department of Computer Science
Virginia Tech
Blacksburg, VA, USA
bmayer@cs.vt.edu

Thomas Mortier
Department of Environment, H-CEL,
Ghent University
Ghent, Belgium
thomasf.mortier@ugent.be

Victor Deklerck
Meise Botanic Garden
Meise, Belgium
victor.deklerck@worldforestid.org

Jakub Truszkowski
Chalmers University of Technology &
World Forest ID
Gothenburg, Sweden
jakub.truszkowski@worldforestid.org

John C. Simeone
Simeone Consulting, LLC
Littleton, New Hampshire, USA
simeoneconsulting@gmail.com

Marigold Norman
World Forest ID
Washington, DC, USA
marigoldwalkins@gmail.com

Jade Saunders
World Forest ID
Washington, DC, USA
jade.saunders@runbox.com

Chang-Tien Lu
Department of Computer Science
Virginia Tech
Alexandria, VA, USA
clu@vt.edu

Naren Ramakrishnan
Department of Computer Science
Virginia Tech
Alexandria, VA, USA
naren@cs.vt.edu

Abstract

Illegal logging poses a significant threat to global biodiversity, climate stability, and depresses international prices for legal wood harvesting and responsible forest products trade, affecting livelihoods and communities across the globe. Stable isotope ratio analysis (SIRA) is rapidly becoming an important tool for determining the harvest location of traded, organic, products. The spatial pattern in stable isotope ratio values depends on factors such as atmospheric and environmental conditions and can thus be used for geographic origin identification. We present here the results of a deployed machine learning pipeline where we leverage both isotope values and atmospheric variables to determine timber harvest location. Additionally, the pipeline incorporates uncertainty estimation to facilitate the interpretation of harvest location determination for analysts. We present our experiments on a collection of oak (*Quercus* spp.) tree samples from its global range. Our pipeline outperforms comparable state-of-the-art models determining geographic harvest origin of commercially traded wood products, and has been used by European enforcement agencies to identify harvest location misrepresentation. We also identify opportunities for further advancement of our framework and how it can be generalized to help identify the origin of falsely labeled organic products throughout the supply chain.

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

KDD '25, Toronto, ON, Canada

© 2025 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-1454-2/2025/08
https://doi.org/10.1145/3711896.3737201

CCS Concepts

• **Theory of computation** → **Gaussian processes**; • **Computing methodologies** → *Multi-task learning*; *Uncertainty quantification*; • **Applied computing** → *Physical sciences and engineering*; *Computer forensics.*

Keywords

Stable isotope ratio analysis (SIRA), Gaussian processes, Uncertainty estimation, Multitask learning, ML Applications

ACM Reference Format:
Shailik Sarkar, Raquib Bin Yousuf, Linhan Wang, Brian Mayer, Thomas Mortier, Victor Deklerck, Jakub Truszkowski, John C. Simeone, Marigold Norman, Jade Saunders, Chang-Tien Lu, and Naren Ramakrishnan. 2025. Chasing the Timber Trail: Machine Learning to Reveal Harvest Location Misrepresentation. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2 (KDD '25), August 3–7, 2025, Toronto, ON, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3711896.3737201

1 Introduction

Illegal logging is a global problem with serious economic, social, and ecological impacts. It is the most profitable natural resource

crime with an estimated annual value between $52 billion and $157 billion, as well as the third largest transnational crime, after counterfeiting and drug trafficking [9]. Illegal logging causes destruction to forest ecosystems, the local flora and fauna, and the people whose livelihoods depend upon forests for goods and services. The effects of illegal logging are not only felt in the area where the illicit activity takes place but also impact associated trade and international markets. Two decades ago it was determined that the value of US-harvested hardwoods (e.g., maple, oak, ash, poplar) had been depressed by 7-16% due to the amount of illegal timber in global trade [19]. Given the complexity of global supply chains and the environmental and societal harm that practices such as illegal resource harvesting present, it is increasingly difficult for stakeholders (e.g., consumers, businesses involved in international trade, and governments seeking to enforce regulations) to ensure that products do not contribute to such environmental and social ills.

Knowing whether illegal logging has taken place starts with identifying where and when the specific trees in question were harvested, and continues to other key data elements along the supply chain. The complexity of global supply chains and the lack of key data passing through the supply chain makes it near impossible to track and trace a product from consumer back to its harvest origin. Supply chain actors may try to misrepresent the harvest location of timber products for a variety of reasons, some of which may be deliberate (e.g., illegal logging or other type of illicit or fraudulent activity such as sanctions evasion), or honest mistakes in the reported geographic harvest origin (e.g., at a fine scale spatial resolution, misrepresentation due to inaccuracies in GPS signal). Examples of deliberate misrepresentation of harvest location include:

- In 2015, US flooring importer Lumber Liquidators plead guilty to environmental crimes related to importing illegally harvested Russian wood, which included felony misdeclarations on its US import documentation, claiming the species of oak used was harvested in Germany, even though that specific species only grows in Northeast Asia [17];
- Due to Russia's 2022 war in Ukraine and the sanctions regime put in place by many countries, there has been increased scrutiny on sanctions evasion of Russian timber entering Europe by misdeclaring the harvest origin [5, 22].

What is now possible, due to global efforts including the effort described here, is to use stable isotope ratio analysis (SIRA) to validate the claimed geographic harvest origins from finished forest products such as furniture or flooring. Stable isotopes are chemical variants of elements that do not go through radioactive decay. By measuring stable isotope values through mass spectrometry [1] and comparing it to a location-referenced database, we can help understand the origin of organic products. Previous work has shown how it can be used to trace the origin of items such as wood, seafood, agricultural products, and cotton [4, 12, 14, 23, 25, 28].

We focus on wood and forest products, which play a critical role in the fight against climate change and, as mentioned earlier, the illegal harvesting of trees is the most profitable natural resource crime [9, 13]. The use of stable isotope testing as a tool to determine the origin of such products is seen as a critical mechanism to identify instances of false or misrepresented harvest location claims of wood and forest products, thus assisting in uncovering instances of illegal logging and sanctions evasion.

Our key contributions are:

(1) We present a comprehensive multi-task Gaussian process modeling framework that supports the incorporation of auxiliary data such as climate layers to support origin determination. This enables the incorporation of environmental factors, imputing uncertainty to predictions, and multimodal feature integration.
(2) Our work is a **deployed** machine learning pipeline [16] wherein physical samples are collected, subject to tests, and injected into our model to help European enforcement agencies in combating illegal timber trade by demonstrating that a claimed harvest location other than Russia is not viable.
(3) While (due to confidentiality reasons) we are unable to showcase specific instances of mislabeled data in this paper, we demonstrate accuracy profiles of our approach in a controlled experiment that illustrates the interplay between SIRA values and atmospheric variables and how they affect our ability to reveal harvest location misrepresentation. This goes beyond traditional ML pipelines that only predict isotope values into an end-to-end approach that supports decision-making by enforcement agencies.

2 Methodology

2.1 Stable Isotope Ratios

A stable isotope ratio refers to the ratio of two stable isotopes of a single element [3]. The natural variation observed for this ratio is determined by underlying mechanisms that are affected by a range of different factors including but not limited to environmental, atmospheric, soil, metabolic fraction, or other characteristics specific to a species [20, 24, 26]. The measurement of stable isotope ratios is commonly expressed using delta (δ) notation, which represents variations in the isotope ratio in parts per thousand (‰, per mil). The most commonly analyzed stable isotopes include those of oxygen, carbon, hydrogen, nitrogen, and sulfur, which are widely used in various scientific fields such as geology, ecology, and climate science. For each element, the delta notation indicates the deviation of the sample's isotope ratio from a standardized reference. In general, it is defined as follows:

$$\delta^{n}\mathrm{E} = \left(\frac{\left(\frac{^{n}\mathrm{E}}{^{n'}\mathrm{E}}\right)_{\text{sample}}}{\left(\frac{^{n}\mathrm{E}}{^{n'}\mathrm{E}}\right)_{\text{standard}}} - 1 \right) \times 1000$$

where E represents the element of interest, ^{n}E is the heavier isotope, and $^{n'}$E is the lighter isotope. The following are the four stable isotope ratios used in this paper:

- δ^{13}C: a measure of the ratio of two stable isotopes of Carbon (^{13}C and ^{12}C).
- δ^{18}O: a measure of the ratio of two stable isotopes of Oxygen (^{18}O and ^{16}O).
- δ^{2}H: a measure of the ratio of two stable isotopes of Hydrogen (^{2}H and ^{1}H).

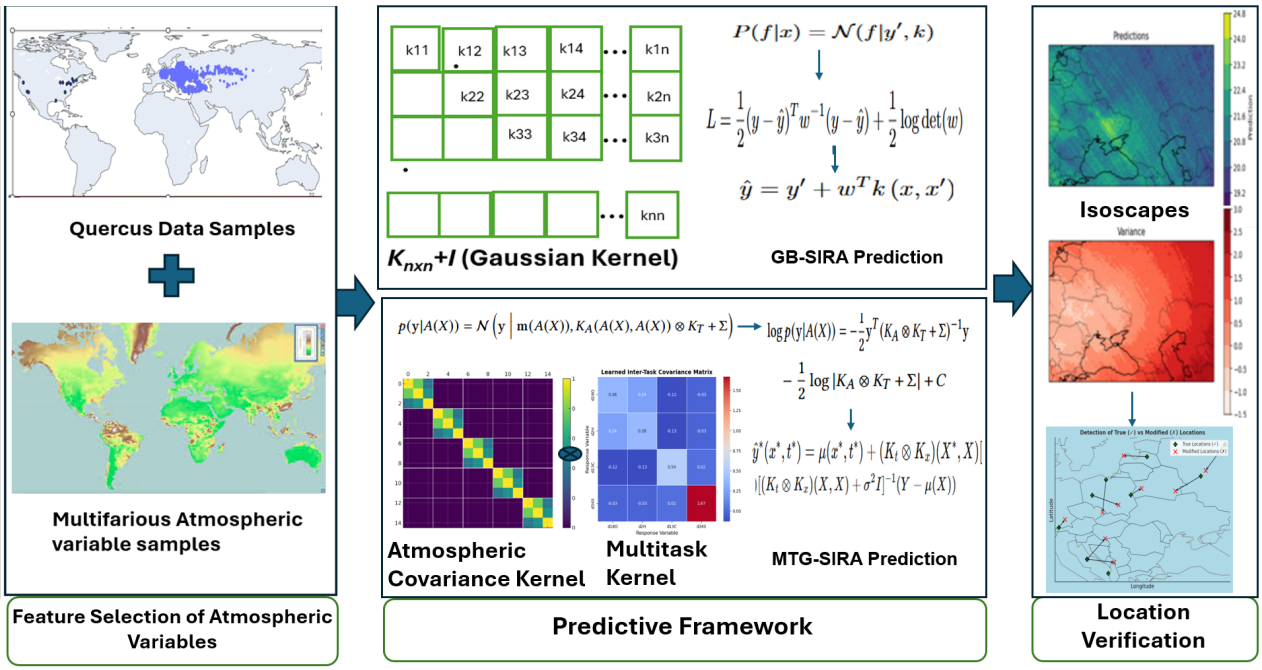

Figure 1: ML Pipeline for Verifying Harvest Location Claims using SIRA.

- δ^{34}S: a measure of the ratio of two stable isotopes of Sulfur (^{34}S and ^{32}S).

In this paper, we propose an end-to-end ML Pipeline that leverages different techniques for different stages of the workflow as illustrated by Figure 1. The symbols used in the methodology are listed in Table 1 for better comprehension.

2.2 Atmospheric Data Layers

Stable isotope ratios are largely dependent on various environmental factors, including but not limited to precipitation, water vapor pressure, and reflected shortwave radiation [6, 24]. The ecological process that facilitates enrichment, or the lack thereof, for stable isotopes in oak trees is a lengthy process. This process can be influenced by climatic patterns observed over long periods. To capture overall climatic conditions, including regularities and anomalies, we collect comprehensive data on 25 different atmospheric properties [2, 8, 15] spanning over 20 years across the world.

2.3 Modeling Framework

Gaussian process methods are adept at modeling spatial data but to capture complex non-linear relationships with the atmospheric variables, decision tree-based boosting algorithms have been shown to be effective. We describe a combined Gaussian process-mixed effect modeling approach motivated by [21]. Given a set of locations $X = \{x_1, x_2..., x_n\}$ with corresponding atmospheric features $A(X)$ = $\{A(x_1), A(x_1), ..., A(x_n)\}$, where $A(x_i)$ is the atmospheric feature at location x_i, we first aim to configure two separate learners, as outlined next. We first provide a description of the mixed-effects learner, followed by an explanation of Gaussian Process Regression (GPR).

To model the non-linear relationship between atmospheric variables and stable isotope ratios, we use mixed-effect modeling comprised of multiple tree-based learners. The algorithm tries to learn a set S of tree-based base learners $f(\cdot)$ in a function space H, defined as the linear function span of S. Hence, the prediction derived from this mixed-effect model is defined as $Y' = F(A(X))$.

GPR is particularly popular in this context as it naturally aids in uncertainty estimation. The observed values are viewed as the realization of a random process defined as $P(f|X) = \mathcal{N}(f|\mu, K)$ where $\mu = [m(x_1), m(x_2)...m(x_n)]$ is the mean estimate and K is the parameterized covariance matrix where K_{ij} is the kernel covariance function of x_i and x_j. Traditionally, the Gaussian process is assumed to be zero mean. However, we combine the output of the base decision tree learner by modifying the probability density function in the following way:

$$P(f|X) = \mathcal{N}(f|Y', K) \tag{1}$$

In this work, the Gaussian process takes an input $X = \{x_1, x_2, x_3...x_n\}$ where x_i denotes the location of ith sample in the training set. The target variable Y is a single isotope ratio. GPR fits the following kernel coefficient:

$$w = [K_{nn} + \Sigma]^{-1} Y \tag{2}$$

where Σ is the error term that corresponds to the uncertainty estimation and K_{nn} is the covariance matrix of dimension $(n \times n)$ where n is the number of samples in the training set.In our case, K_{nn} is the kernel function of our choosing. We use a combination of radial basis function, periodic, and rational quadratic function as the covariance function. Therefore the learned covariance matrix can be defined as $K = w^T[K_{nn} + \Sigma]$

Table 1: Summary of Symbols used in our framework.

Symbol	Meaning	Symbol	Meaning
δ^nE	Delta isotope ratio notation	^{n}E	Heavier stable isotope
$^{n'}$E	Lighter stable isotope	$X = \{x_i\}$	Sample location set
$A(x)$	Atmospheric vector at x	$A(X)$	Atmospheric data for X
Y	Observed isotope values	Y'	Prediction from $F(A(X))$
$F(A(X))$	Mixed-effect learner	$f(\cdot)$	Base tree learner
$P(f\|X)$	Posterior over GP function	μ	GP mean vector
K	GP covariance matrix	K_{ij}	Covariance between x_i and x_j
K_{nn}	Covariance of training data	Σ	Noise/error matrix
w	GP weight vector	X^*	Test locations
$\hat{Y}$	Final prediction	L	GP loss (log-likelihood)
$R(F, w)$	Joint risk functional	$\mathcal{H}$	Space of base learners
Θ	Parameter space for w	$f_{T_i}(A(x))$	Output for task T_i
K_m	Multitask kernel	K_T	Task covariance matrix
K_A	Feature covariance matrix	λ_1, λ_2	Kernel weight scalars
$K_{\text{RBF}}, K_{\text{Matérn}}$	RBF / Matérn kernels	$\ell_{T_i,d}$	ARD lengthscale for task T_i, feature d
$\text{FI}_{T_i}(d)$	Feature importance for T_i	$\mathbf{L}$	Cholesky factor of K_T
Y_{true}	Claimed isotope values	x_{mod}	Claimed origin location
μ_{mod}	GP mean at x_{mod}	$\mathbf{K}_{\text{mod}}$	Covariance matrix at x_{mod}
χ^2	Chi-squared statistic	p_{mod}	P-value of hypothesis test
α	Significance level	–	–

Now, for a new set of unseen samples X^*, the prediction is calculated as:

$$\hat{Y} = Y' + w^T K\left(X, X^*\right) \tag{3}$$

2.4 Joint Optimization

As described in [21] we aim to combine the two learners through a joint optimization goal. Here the loss function is the negative log-likelihood function as described below:

$$L = \frac{1}{2}(Y - \hat{Y})^T K^{-1}(Y - \hat{Y}) + \frac{1}{2}\log\det(K) \tag{4}$$

If $\mathbf{w}$ is the parameter of the Gaussian Process co-efficient kernel K and $F(\cdot)$ is the mixed-effect learner $F(A(X))$, the goal is to find the joint minimizer:

$$(\hat{F}(\cdot), \hat{\mathbf{w}}) = \underset{(F(\cdot),\mathbf{w})\in(\mathcal{H},\Theta)}{\operatorname{argmin}} R(F(\cdot), \mathbf{w}) \tag{5}$$

Here, $\mathcal{H}$ is the superset of all base learners and Θ is the parameter space for w. $R(F(\cdot), \mathbf{w})$ is a risk functional defined as:

$$R(F(\cdot), \mathbf{w}) : \quad (F(\cdot), \mathbf{w}) \mapsto L \tag{6}$$

Here, L is the loss function described earlier. R is determined by evaluating $F(A(X))$, $\hat{Y}$ and then calculating the loss function L. The risk factor R is minimized using the Gaussian Process boosting algorithm. First, for iteration i we determine w_i as the following:

$$\mathbf{w}_i = \operatorname{argmin}_{\mathbf{w}\in\Theta} L\left(y, F_{i-1}, \mathbf{w}\right) \tag{7}$$

where L is the loss function denoted in equation 5. Then, we update base learner F_i through a Newton step:

$$F_i = F_{i-1} - \frac{R'\left(F_{i-1}\right)}{R''\left(F_{i-1}\right)} \tag{8}$$

where R' and R'' denote first and second-order derivatives of the reward function for the $(i-1)^{\text{th}}$ iteration of F.

2.5 Modeling Dependencies between Isotopes

Although multivariate Gaussian Processes in the form of co-kriging have been explored for stable isotope prediction, the idea of inter-task dependencies has not been explored yet and is novel to our approach. Due to the strong correlations that stable isotope ratios often exhibit, we hypothesize that using a Gaussian process to leverage both the spatial correlation of atmospheric variables and the inter-task dependencies of different stable isotopes could greatly benefit SIRA prediction. Further, we aim to leverage automatic relevance judgments from amongst the multivariate input of atmospheric variables. Similar to the GPR model, the Gaussian Process for the multi-task model defines a distribution over the function f defined over the space spanned by atmospheric vector A(X):

$$\begin{aligned}\mathbf{f}(A(X)) &= \left[f_{T_1}(A(X)), \ldots, f_{T_M}(A(X))\right]^T \\ &\sim \mathcal{GP}\left(m_m(A(X)), K_m\right)\end{aligned} \tag{9}$$

where $f_{T_i}(A(x))$ can be viewed as the objective function for task i, with the task corresponding to one of the four stable isotope ratio being modeled. $m_m(A(X))$ is the mean function for the atmospheric features of locations X and K_m is the multitask covariance kernel. Given two locations x and x', the covariance between the feature vector A(x) and A(x') across multiple tasks is given by a Kronecker product structure:

$$k_{MT}((A(x), T_i), (A(x'), T_j)) = K(A(x), A(x')) \otimes K_{T_i,T_j} \tag{10}$$

where $K(A(x), A(x'))$ represents the covariance function for the atmospheric vector corresponding to location x and x' and K_{T_i,T_j} is the task covariance matrix between two tasks T_i and T_j.

2.6 Interpretability via the Covariance Matrix

The feature covariance matrix, i.e., K(A(x), A(x')), is parameterized by automatic relevance determination of the length scale for the chosen covariance matrix. As Matern kernels have been proven to be effective in leveraging the sudden change in the latent feature

space, and an RBF kernel is effective in capturing a smoother correlation in the feature space, we use a combination of these two kernels to constitute our covariance matrix:

$$K(A(x), A(x')) = \lambda_1 K_{\text{RBF}}(A(x), A(x')) + \lambda_2 K_{\text{Matérn}}(A(x), A(x')) \tag{11}$$

$$K_{\text{RBF}}(A(x), A(x')) = \sigma^2 \exp\left(-\frac{1}{2}\sum_{d=1}^{D}\frac{(A_d(x) - A_d(x'))^2}{\ell^2_{T_{i,d}}}\right) \tag{12}$$

$$K_{\text{Matérn}}(A(x), A(x')) = \sigma^2 \prod_{d=1}^{D}\left(1 + \frac{\sqrt{3}|A_d(x) - A_d(x')|}{\ell_{T_{i,d}}}\right) \exp\left(-\frac{\sqrt{3}|A_d(x) - A_d(x')|}{\ell_{T_{i,d}}}\right) \tag{13}$$

In the above, σ^2 is the variance, $\ell_{T_{i,d}}$ is the ARD lengthscale for task i and atmospheric variable at dimension d, and $|A(x) - A(x')|$ denotes the Euclidean distance between the feature vector at location x and x'. Through optimizing the objective function we learn the ARD lengthscale value for each dimension of the feature space which is used to extract feature importance:

$$\text{Feature Importance } T_i(d) = \frac{1}{\ell_{T_{i,d}}} \tag{14}$$

In addition to feature-based interpretability, we also leverage task-specific insights through factorization of the task covariance matrix K_T:

$$\mathbf{K}_T = \mathbf{L}\mathbf{L}^T \tag{15}$$

where L is a lower triangular matrix.

2.7 Multitask Loss Function

Unlike traditional GPR we formulate a joint likelihood function over all the response variables defined as:

$$p(\mathbf{Y}|A(X)) = \mathcal{N}\left(\mathbf{Y} \,\middle|\, \mathbf{m}(A(X)), K_A(A(X), A(X)) \otimes K_T + \Sigma\right) \tag{16}$$

and the marginal log-likelihood derived as:

$$\log p(\mathbf{Y}|A(X)) = -\frac{1}{2}\mathbf{Y}^T(K_A \otimes K_T + \Sigma)^{-1}\mathbf{Y} - \frac{1}{2}\log|K_A \otimes K_T + \Sigma| + C \tag{17}$$

where C is a normalization constant.

2.8 Location Verification

Given Y_{true}, the given value of the isotope ratio, our objective is to determine whether the claim about the origin being x_{mod} is true or not. First, we compute the likelihood of observing Y_{true} for x_{mod} :

$$\log p(Y_{\text{true}}|A_{\text{mod}}) = -\frac{1}{2}(Y_{\text{true}} - \mu_{\text{mod}})^T\mathbf{K}^{-1}_{\text{mod}}(Y_{\text{true}} - \mu_{\text{mod}}) - \frac{1}{2}\log|\mathbf{K}_{\text{mod}}| - \frac{M}{2}\log 2\pi \tag{18}$$

Here, M is the number of isotope ratios or length of the Y_{true} vector which is also the degrees of freedom for the subsequent hypothesis test; μ_{mod} is the predicted mean for the location x_{mod} given atmospheric variable A_{mod}; and $\mathbf{K}_{\text{mod}}$ is the corresponding covariance matrix.

We can now determine the veracity of the claim by the following hypothesis test:

- H_0: The test sample Y_{true} comes from the location x_{mod}
- H_1: The test sample Y_{true} does not come from the location x_{mod}

Leveraging the trained GP model we formulate:

$$\chi^2 = (Y_{\text{true}} - \mu_{\text{mod}})^T\mathbf{K}^{-1}_{\text{mod}}(Y_{\text{true}} - \mu_{\text{mod}}) \tag{19}$$

According to the null hypothesis, χ^2 follows a chi-squared distribution with M degrees of freedom which is the number of stable isotope ratios being modeled. The p-value can be expressed as :

$$p_{\text{mod}} = P(\chi^2 > \chi^2_{\text{obs}}) \tag{20}$$

Finally, the decision rule for classifying a claim as false is upper bound by a significance threshold α:

$$\text{Decision:} \quad p_{\text{mod}} < \alpha \Rightarrow \text{Location Claim False} \tag{21}$$

3 Experimental Results

3.1 Datasets

As described earlier, while our system is deployed, we are unable to present results due to confidentiality constraints. We describe results on a different dataset gathered from samples of the genus *Quercus* from countries such as China, the United States, Ukraine, and Russia. Stable isotope ratio measurements for each sample were calculated and aggregated as described in [27]. Each entry contained stable isotope ratio measurements of oxygen, hydrogen, sulfur, and carbon (along with their GPS coordinates). Our atmospheric data includes isotopic composition of precipitation, water vapour, shortwave radiation, temperature, and many more factors [2, 8, 15]. In overall, we use 25 atmospheric variables for 20 years. In mapping the atmospheric variables to the sampled location of the *Quercus* spp. values, we found six of the atmospheric variables to have more than 50% missing values due to the sparse nature of the data; these variables are discarded. This retains 19 atmospheric variables, each having 12 months of data for every year, for 20 years.

3.2 Research Questions

Our goals in this study are to characterize the effectiveness of our methodology for modeling stable isotope values and subsequent location claim verification:

- **RQ1:** How does the proposed framework compare to existing approaches to stable isotope ratio prediction?
- **RQ2:** How do different architectural choices (such as, combining tree boosting with Gaussian processes, multivariate GP modeling, or modeling inter-task correlation) benefit the outcome compared with using any of these techniques individually?
- **RQ3:** Does the incorporation of atmospheric variables and the feature selection methodology contribute to more accurate forecasts?

- **RQ4:** What implicit explainability can the architecture provide for atmospheric variables and inter-task correlation?
- **RQ5:** How effective is the proposed framework at successfully trapping false claims of timber origin?

3.3 Implementation Details

For all isotope ratios we split the dataset into an 80:20 training/test split. The splits are stratified by geography to ensure generalization across regions. We perform K-fold cross-validation with $k = 5$ and choose the average score on all evaluation metrics for reporting. For the GP-Boost algorithm of GB-SIRA, we perform parameter searching and finalize the learning rate to be 0.03, epochs at 250 and max depth to be 5. For the multitask-learning GP model, we fix the learning rate to be 0.01 and the number of epochs to 500. Training a full GB-SIRA model on 800+ samples completes in less than 2 minutes on a CPU with moderate specs (16-core, 64 GB RAM), and inference per sample takes 0.2 seconds. Training the MTG-SIRA model on the same number of samples across 4 isotope tasks is completed within 3 minutes on a single GPU. Inference time per sample remains in the range of 0.2-0.5 second per task when batched. The model is evaluated using standard metrics for regression tasks, R^2, and Root mean square error (RMSE). The R^2 value is a widely used metric for evaluating regression problems that indicates the proportion of variance in dependent variables that can be directly explained by the independent variables. RMSE helps understand how close the prediction is to the ground truth across the test samples.

3.4 RQ1: Baseline Experiments

We compare our methodology (MTG-SIRA to denote the full multi-task framework; and GB-SIRA to denote our approach minus the multi-task formulation) against the following alternative approaches:

- Watkinson et al. [27]: This work proposes the use of ordinary kriging for spatial interpolation of the isotope ratio values. They were used to produce predictions for carbon, oxygen, hydrogen, and sulfur stable isotope ratios on *Quercus* spp., similar to our problem statement but only on samples from the United States. However, we use their methodology on our data for comparison.
- Watkinson et al. [28]: This work uses Universal Kriging to predict the isotope ratio values of timber in Solomon Islands.
- Truszkowski et al. [23]: This work, similar to ours, uses atmospheric data for stable isotope ratio prediction but the covariance matrix is modified by using location-specific atmospheric variables making it analogous to a co-kriging method.
- Support vector regression (SVR): This is a regression model variation of SVM that has shown to be effective extensively for a dataset of this size. This approach also serves as a traditional ML baseline.

The experiments as detailed in Table 2 demonstrate that, in terms of R^2, our proposed frameworks GB-SIRA and MTG-SIRA outperform other approaches for 2 out of the 4 isotope ratio prediction tasks. However, even for δ^{13}C and δ^{34}S, the difference with respect to the best-performing model is negligible. For isotopes such as δ^{2}H and δ^{18}O, there is significant improvement witnessed.

In terms of RMSE, we also observe the lowest RMSE value for MTG-SIRA for all stable isotope ratios except δ^{34}S for which GB-SIRA performs the best. We note that Watkinson et al. [27] did not report RMSE. The performance of both GB-SIRA and MTG-SIRA overall compares favorably, demonstrating the effectiveness of the proposed ML framework. In general, the performance for δ^{34}S and δ^{13}C is not high across the models, a result in line with past related work. Finally, the overall performance improvement on the multitask model MTG-SIRA indicates the effectiveness of modeling inter-task dependency.

3.5 RQ2: Architectural Choices

To answer this research question we ablate each individual component after the feature selection module to compare against the final model. The result of this experimental setting is described in Table 3. The framework without the Gaussian Process element becomes a decision-tree boosting algorithm called lightGBM which takes as input the output of the feature selection module and doesn't model the spatial element. On the other hand, without the boosting algorithm combined, GB-SIRA only takes the sample longitude and latitude as input which makes it the traditional GPR. Multivariate GP is the variation of the model that models spatial correlation by using the atmospheric variable as latent feature space. For the boosting algorithm, we observe encouraging performance on $\delta^{34}S$ and $\delta^{13}C$. However, when combined with Gaussian processes, the performance invariably increases. Similarly, even though multivariate Gaussian process modeling shows encouraging performance on $\delta^{18}O$, the approach of incorporating a multitask-learning module does improve the overall performance of the model. Hence, given the importance of task-specific interpretability and uncertainty estimation, MTG-SIRA proves to be the best iteration of our proposed framework. However, this is a decision to be made in consultation with all stakeholders.

3.6 RQ3: Role of Atmospheric Variables

To address this question, we look at the two best-performing models from Table 2. Since both these models take advantage of atmospheric variables, we create an experimental setting that facilitates both the inclusion and exclusion of the feature selection module. The result as described in Table 4 shows that the inclusion of the feature selection step before modeling produces a better prediction for both models. Furthermore, we also show the experimental results with and without the atmospheric variables in Table 2. There is a noticeable improvement for the prediction of oxygen and hydrogen stable isotope ratios.

3.7 RQ4: Explainability

As discussed in Equations 2.3 and 2.6, feature importance is inversely proportional to learned lengthscales corresponding to each variable. Therefore, the trained model for MTG-SIRA can be used to visualize the feature importance of atmospheric variables for stable isotope prediction. As shown in Fig 2, the model is able to identify variables such as "cloud water content", "water vapour", "land surface temperature" and "precipitation" as some of the most important features. This is consistent with existing literature [6] on the subject that finds strong correlation between certain stable

Table 2: Baseline comparison of predictive models based on R^2 and RMSE for SIRA values.

Baseline Comparison	Atmospheric Variables	R^2				RMSE			
		δ^{18}O	δ^{13}C	δ^{2}H	δ^{34}S	δ^{18}O	δ^{13}C	δ^{2}H	δ^{34}S
Watkinson et al. [27]	–	0.470	0.320	0.700	**0.690**	–	–	–	–
Watkinson et al. [28]	–	0.856	0.301	0.730	0.601	0.673	0.779	6.112	1.233
Truszkowski et al. [23]	X	0.869	**0.331**	0.790	0.667	0.631	0.757	6.279	1.070
SVR	X	0.754	0.313	0.750	0.682	0.940	0.773	6.410	1.040
GB-SIRA	X	0.878	0.322	0.840	0.689	0.742	0.768	5.911	**1.011**
MTG-SIRA	X	**0.899**	0.310	**0.858**	0.673	**0.687**	**0.679**	**5.201**	1.029

Table 3: Ablation study for various model choices evaluated using R^2 and RMSE.

Model	δ^{18}O	δ^{13}C	δ^{2}H	δ^{34}S
R^2				
Boosting	0.865	**0.329**	0.819	0.675
GPR	0.856	0.301	0.730	0.601
Boosting+GPR	0.878	0.322	0.840	**0.689**
Multivar GP	0.880	0.299	0.834	0.625
Multi-Task GP	**0.899**	0.310	**0.858**	0.673
RMSE				
Boosting	0.801	0.780	5.820	1.070
GPR	0.773	0.779	6.112	1.233
Boosting+GPR	0.742	0.768	5.911	**1.011**
Multivar GP	0.691	0.725	5.890	1.033
Multi-Task GP	**0.687**	**0.679**	**5.201**	1.029

isotope values and the aforementioned atmospheric variables. The model's ability to identify such factors reiterates the effectiveness of MTG-SIRA as a self-explainable mechanism for feature-level interpretation. Similarly, we aim to factorize the inter-task dependency and use the lower triangular matrix to gain insights into how different elemental stable isotope ratios are correlated with each other. Figure 3 shows strong indication of a positive correlation between $\delta^{18}O$ and $\delta^{2}H$ and a negative correlation between $\delta^{34}S$ and other stable isotopes. This can inform future usage of stable isotope measurements for location determination.

3.8 RQ5: Evaluating Claims of Timber Origin

As mentioned earlier, our approach is in use by European enforcement agencies. However, because we cannot showcase results on real data, we formulate a controlled experiment wherein we modify the true (ground truth) location of a sample by a specified distance d. Thus x is modified, resulting in a modified atmospheric vector for x. This can be formulated as:

$$x_{\text{mod}} = x + \Delta x, \quad \Delta x \sim \mathcal{N}(0, \sigma^2_{\text{perturb}}) = d \tag{22}$$

$$A_{\text{mod}} = A(x_{\text{mod}}) \tag{23}$$

Given Y_{true} as the true value of the isotope ratio, our objective is to determine whether the claim about the origin being x_{mod} is true or not. Then we calculate the accuracy in identifying the false claims for each value of d ranging from 500 to 5000 km in geodesic distance.

Fig 4 showcases the results. When the altered distance is within 500km, the success rate is not high (approximately 60%), consistent with the fact that atmospheric variables likely do not shift significantly unless elevation is accounted for. From this point, we begin to see a gradual improvement in false claim detection as the distance d increases up until 2500 km, reaching an accuracy of over 80%. As the density of training samples is high in specific areas, when the modified location falls outside the range of low uncertainty estimates for the model, we see a steady decline in the accuracy of a verification claim, ultimately becoming stable after 4000 km. This is a significant finding that can guide authorities to the responsible use of our approach for claim verification.

4 Description of Deployment

In practice, the system has two use-cases:

First, for a given location, the system outputs: i) the prediction of isotope ratios, ii) a confidence interval based on predictive uncertainty, and iii) feature-level interpretability via ARD kernel length scales (e.g., precipitation, temperature, vapor pressure etc. as key drivers).

Second, for a given sample with measured stable isotope ratio values and a claim of the geographic harvest origin of the sample, the system outputs: i) the probability of the sample originated from the claimed harvest location, ii) a binary flag indicating whether the location claim is true or false.

The system also enables triaging: low-confidence, high-risk predictions are escalated for manual reliability. Furthermore, locations with low-confidence and high-risk predictions are also noted for driving the future collection of samples, assisting a more effective approach in sample collection.

Our framework is in use by European enforcement agencies to assist in demonstrating that a claimed harvest location other than

Table 4: Comparison of MTG-SIRA and GB-SIRA with and without Feature Selection.

Method	R^2				RMSE			
	δ^{18}O	δ^{13}C	δ^{2}H	δ^{34}S	δ^{18}O	δ^{13}C	δ^{2}H	δ^{34}S
MTG-SIRA without Feature Selection	0.860	0.271	0.816	0.632	0.683	0.801	6.350	1.210
GB-SIRA without Feature Selection	0.874	0.289	0.781	0.655	0.662	0.813	6.210	1.194
MTG-SIRA with Feature Selection	**0.899**	0.310	**0.858**	0.673	**0.687**	**0.679**	**5.201**	1.029
GB-SIRA with Feature Selection	0.878	**0.322**	0.840	0.689	**0.742**	**0.768**	**5.911**	**1.011**

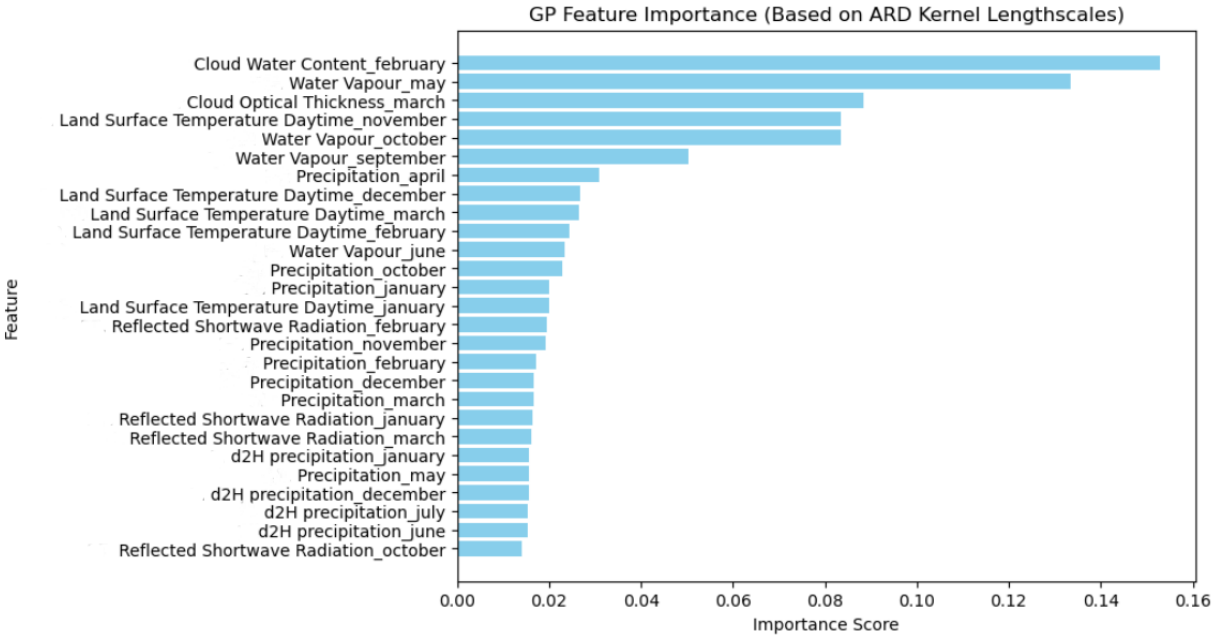

Figure 2: Atmospheric Feature Importance.

Russia is not viable. See also reported coverage of our work [16]. Due to confidentiality reasons, we present results here on a different species (Oak; *Quercus* spp.). For instance, see isoscape predictions around the Russia/Ukraine border region in Fig. 5. In one ongoing analysis, we are testing 59 samples from corporate partners, operating primarily in the UK and USA. Results indicate a 41% fail rate for harvest claims; in other words, 41% of importer claims tested were incorrect.

5 Conclusion

We have presented a novel ML framework that uses SIRA to determine the origin of organic products. It serves as a holistic end-to-end predictive pipeline that can facilitate feature selection and systematic analysis of recorded and predicted isotope ratio values. Our approach to multimodal feature integration also enables us to leverage the vast information present in atmospheric data by systematic feature selection methods. Further, such incorporation of atmospheric variables into the methodology supports interpretability that prior methods do not support.

A key real-world implication of our methodology is that it can create species-wide isoscapes that predict sub-national variability even in areas without ground truth sample data. This is particularly useful if the security context makes sample data collection challenging or impossible. Furthermore, the actionability of these species' isoscapes can be improved by visualizing uncertainty estimations to communicate the relative confidence of a predicted value in a particular region. This enables real-world decision-making based on predictions with low uncertainty values to manage legal and financial liabilities arising from timber supply contracts and law enforcement activities.

Using our methodology, being able to infer the origins of an unseen sample with varying degrees of confidence has proven to be an effective tool used to enforce timber trade regulations and sanctions, such as the sanction on importing Russian and Belarusian timber in the EU [14, 23].

There is promise for additional social and environmental impact of this work, given that the application of our SIRA methods can be applied to origin testing for a variety of organic products. Our work

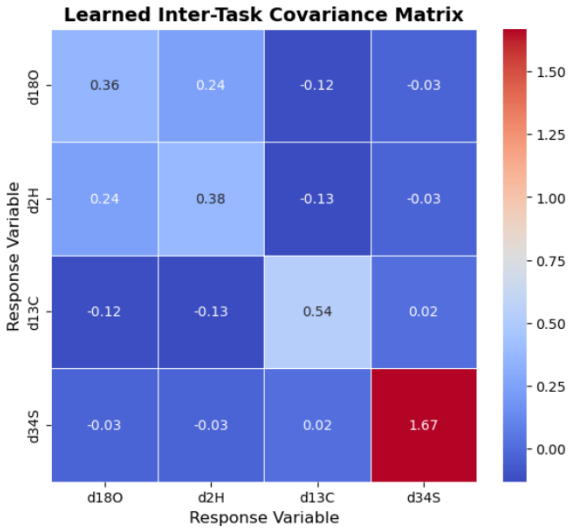

Figure 3: Inter-task Dependency.

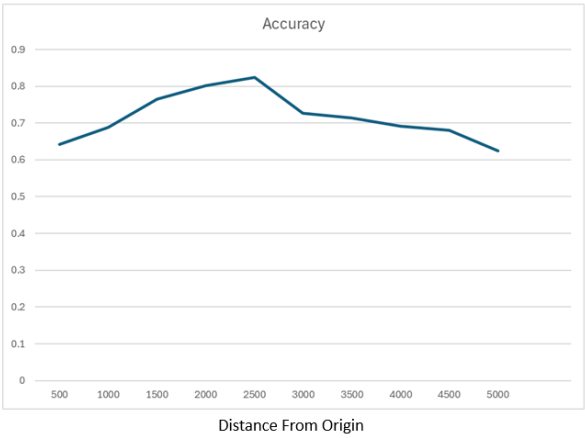

Figure 4: Success rate of identifying false claim vs distance from true origin.

is particularly relevant given the EU's Regulation on Deforestation-free Products (EUDR) which requires proving that products made from forest-risk commodities, e.g., cattle, wood, cocoa, soy, coffee, palm oil, and rubber, do not originate from recently deforested land or have contributed to forest degradation [7].

Furthermore, there has been increased scrutiny on the use of forced labor to produce products and food that enter the global supply chain. Products such as garments and apparel made from cotton produced in the Xinjiang region of China by Uyghurs, and fish and seafood harvested by forced labor, have particularly complex supply chains [4, 10, 11, 18]. This work contributes to the advancement of accurate scientific methods used to assist the determination of origin claims for all organic products covered in policies such as the EUDR and the US Lacey Act and Uyghur Forced Labor Prevention Act.

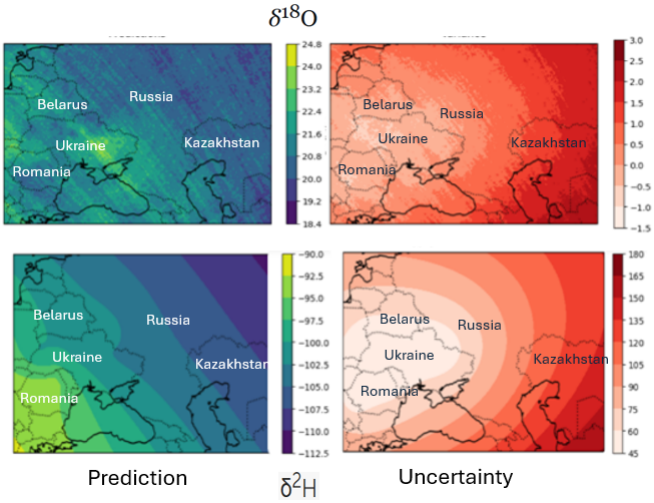

Figure 5: Case Study: Isoscapes for Eastern European Region-Ukraine/Russia Border for Timber.

Acknowledgments

This paper is based upon work supported by the NSF under Grant No. CMMI-2240402. JT is supported by the Swedish Foundation for Strategic Environmental Research MISTRA (Utmana program). 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 sponsors.

References

[1] Andrew Barrie and SJ Prosser. 1996. Automated analysis of light-element stable isotopes by isotope ratio mass spectrometry. *Mass spectrometry of soils. New York, Marcel Dekker* (1996), 1–46.

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

[3] Tyler B Coplen, Carol Kendall, and Jessica Hopple. 1983. Comparison of stable isotope reference samples. *Nature* 302, 5905 (1983), 236–238.

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

[5] Earthsight. 2025. Blood-stained Birch: Exposing the EU trade in Russian conflict ply. https://www.earthsight.org.uk/blood-stained-birch

[6] Ahmed Elbeltagi, Chaitanya B Pande, Manish Kumar, Abebe Debele Tolche, Sudhir Kumar Singh, Akshay Kumar, and Dinesh Kumar Vishwakarma. 2023. Prediction of meteorological drought and standardized precipitation index based on the random forest (RF), random tree (RT), and Gaussian process regression (GPR) models. *Environmental Science and Pollution Research* 30, 15 (2023), 43183–43202.

[7] European Commission. 2023. Regulation on Deforestation-free products. https://environment.ec.europa.eu/topics/forests/deforestation/regulation-deforestation-free-products_en

[8] GJ Huffman, A Behrangi, DT Bolvin, and EJ Nelkin. 2020. GPCP version 3.1 satellite-gauge (SG) combined precipitation data set. *NASA GES DISC: Greenbelt, MD, USA* (2020).

[9] Global Financial Integrity. 2017. Transnational Crime and the Developing World. https://gfintegrity.org/report/transnational-crime-and-the-developing-world/

[10] Katherine Masters. 2023. US Customs finds garments made with banned Chinese cotton. *Reuters* (2023). https://www.reuters.com/markets/commodities/us-customs-finds-garments-made-with-banned-chinese-cotton-documents-2023-09-01/

[11] Arriana McLymore. 2024. Banned Chinese cotton found in 19% of US and global retailers' merchandise, study shows. *Reuters* (2024). https://www.reuters.com/markets/commodities/banned-chinese-cotton-found-19-us-retailers-merchandise-study-shows-2024-05-07/

[12] Wolfram Meier-Augenstein, Helen F. Kemp, Emily R. Schenk, and Jose R. Almirall. 2014. Discrimination of unprocessed cotton on the basis of geographic origin using multi-element stable isotope signatures. *Rapid communications in mass spectrometry: RCM* 28, 5 (March 2014), 545–552. doi:10.1002/rcm.6811

[13] Iana Messetchkova. 2021. Glasgow Leaders' Declaration on Forests and Land Use. https://webarchive.nationalarchives.gov.uk/ukgwa/20230418175226mp_/https://ukcop26.org/glasgow-leaders-declaration-on-forests-and-land-use/

[14] Thomas Mortier, Jakub Truszkowski, Marigold Norman, Markus Boner, Bogdan Buliga, Caspar Chater, Henry Jennings, Jade Saunders, Rosie Sibley, Alexandre Antonelli, et al. 2024. A framework for tracing timber following the Ukraine invasion. *Nature Plants* 10, 3 (2024), 390–401.

[15] NASA. 2024. Reflected Shortwave Radiation neo.gsfc.nasa.gov. https://neo.gsfc.nasa.gov/view.php?datasetId=CERES_SWFLUX_M Accessed 15-08-2024.

[16] Alexander Nazaryan. 2024. New Method That Pinpoints Wood's Origin May Curb Illegal Timber. *The New York Times* (April 2024). https://www.nytimes.com/2024/04/09/science/conflict-timber-russia-europe.html

[17] U.S. Department of Justice. 2015. Lumber Liquidators Inc. Pleads Guilty to Environmental Crimes and Agrees to Pay More Than $13 Million in Fines, Forfeiture and Community Service Payments. *Press Release, Office of Public Affairs* (Oct. 2015). https://www.justice.gov/archives/opa/pr/lumber-liquidators-inc-pleads-guilty-environmental-crimes-and-agrees-pay-more-13-million

[18] PBS News. 2023. Investigation reveals Chinese seafood caught and processed using forced labor sold in U.S. https://www.pbs.org/newshour/show/investigation-reveals-chinese-seafood-caught-and-processed-using-forced-labor-sold-in-u-s Section: World.

[19] Seneca Creek Associates, LLC. 2004. *" Illegal" Logging and Global Wood Markets: The Competitive Impacts on the US Wood Products Industry.* American Forest & Paper Association.

[20] Rolf T. W. Siegwolf, J. Renée Brooks, John Roden, and Matthias Saurer (Eds.). 2022. *Stable Isotopes in Tree Rings: Inferring Physiological, Climatic and Environmental Responses.* Tree Physiology, Vol. 8. Springer International Publishing, Cham. doi:10.1007/978-3-030-92698-4

[21] Fabio Sigrist. 2022. Gaussian process boosting. *Journal of Machine Learning Research* 23, 232 (2022), 1–46.

[22] The Brussels Times. 2024. Over 260 tonnes of Russian timber entered Belgium in breach of EU sanctions. https://www.brusselstimes.com/947366/over-26-tonnes-of-russian-timber-entered-belgium-in-breach-of-eu-sanctions

[23] Jakub Truszkowski, Roi Maor, Raquib Bin Yousuf, Subhodip Biswas, Caspar Chater, Peter Gasson, Scot McQueen, Marigold Norman, Jade Saunders, John Simeone, Naren Ramakrishnan, Alexandre Antonelli, and Victor Deklerck. 2025. A probabilistic approach to estimating timber harvest location. *Ecological Applications* 35, 1 (2025), e3077. doi:10.1002/eap.3077 arXiv:https://esajournals.onlinelibrary.wiley.com/doi/pdf/10.1002/eap.3077

[24] Y Vystavna, I Matiatos, and LI Wassenaar. 2021. Temperature and precipitation effects on the isotopic composition of global precipitation reveal long-term climate dynamics. *Scientific reports* 11, 1 (2021), 18503.

[25] Jishi Wang, Tianjin Chen, Weixing Zhang, Yan Zhao, Shuming Yang, and Ailiang Chen. 2020. Tracing the geographical origin of rice by stable isotopic analyses combined with chemometrics. *Food chemistry* 313 (2020), 126093.

[26] Liuwei Wang, Yuanliang Jin, Dominik J Weiss, Nina J Schleicher, Wolfgang Wilcke, Longhua Wu, Qingjun Guo, Jiubin Chen, David O'Connor, and Deyi Hou. 2021. Possible application of stable isotope compositions for the identification of metal sources in soil. *Journal of Hazardous Materials* 407 (2021), 124812.

[27] Charles J Watkinson, Peter Gasson, Gareth O Rees, and Markus Boner. 2020. The development and use of isoscapes to determine the geographical origin of Quercus spp. in the United States. *Forests* 11, 8 (2020), 862.

[28] Charles J Watkinson, Gareth O Rees, Sabine Hofem, Lina Michely, Peter Gasson, and Markus Boner. 2022. 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), 645222.