GB-SIRA: An ML Framework for Stable Isotope Ratio Analysis to support Environmental Regulation Compliance and Enforcement

Anonymous submission

Abstract

Stable isotope ratio analysis (SIRA) is an increasingly important tool to determine the harvest location of traded, organic, products. The pattern in stable isotope ratio values depends on external factors such as geographic location, atmospheric, or inherent features such as species. Previous work on SIRA for product tracing is either limited by their application of relatively simple statistical analysis and models or does not leverage more powerful machine learning models that can integrate external factors like atmospheric variables. In this work, we implement an end-to-end machine learning framework that not only models the spatial variability and probabilistic distribution of the isotope ratios but also leverages machine learning tools (i.e., feature selection, Gaussian process regression, and boosting algorithms) to incorporate atmospheric variables. Additionally, the pipeline incorporates uncertainty estimation to facilitate the process of origin determination which is useful for tracking illegally shipped items. Our novel predictive pipeline GB-SIRA (Gaussian Process Boosting SIRA) improves the existing benchmark for stable isotope ratio prediction. We present our experiments on a collection of oak (*Quercus* spp.) tree samples from around the world. Our pipeline outperforms comparable state-of-the-art models when tasked with predicting stable isotopes of the oak samples. The deployment of this work has shown early promise in advancing SIRA for harvest location determination which can be used to help enforce social, environmental, and economic trade restrictions by identifying the origin of falsely labeled organic products throughout the supply chain. Furthermore, we propose a plan for additional application and advancement of our framework.

Stable Isotope Ratio Analysis, Machine Learning, Gaussian Process Modeling

Introduction

Stable isotopes are chemical variants of elements that do not go through radioactive decay. The ratio of stable isotopes denotes the relative enrichment of different elemental stable isotopes in a sample which is typically measured by mass spectrometry (Barrie and Prosser 1996) and allows us to understand the enrichment of these isotopes in that sample. 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 (Siegwolf et al. 2022; Wang et al. 2021; Vystavna, Matiatos, and Wassenaar 2021).

Therefore, stable isotope ratios can be useful for understanding the origin of organic products. Previous work has shown how it can be used to trace the origin of items such as timber, seafood, agricultural products, and fiber such as cotton (Truszkowski et al. 2023; Mortier et al. 2024; Watkinson et al. 2022; Cusa et al. 2022; Wang et al. 2020; Meier-Augenstein et al. 2014). Given the complexity of global supply chains and the environmental and societal harm that practices such as illegal resource harvesting and forced labor 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. The use of stable isotope testing as a tool to determine the origin of such products is seen as a critical tool to verify legality and identify instances of false or fraudulent location of harvest claims.

However, many of the works that focus on using SIRA to determine the origin of organic products are limited by:

1. **The lack of a comprehensive machine learning (ML) modeling approach** where methods like regression analysis (Watkinson et al. 2020, 2022) or clustering (Yang, Hutcheon, and Krouse 2001) have been used. These methods, despite being relatively effective for smaller datasets, do not leverage all the information contained in the dataset along with the spatial variation observed in the data. Some of the works in this area try to model the spatial variation of the isoscape but are limited to a more basic probabilistic statistical modeling approach (Mortier et al. 2024) that doesn't facilitate the incorporation of environmental factors like atmospheric variables.
2. **Absence of a holistic end-to-end predictive pipeline that can facilitate feature selection** and a more systematic analysis of predicted isotope ratio values. While the advantage of statistical or probabilistic modeling in estimating uncertainty makes it an attractive choice for the scientific community as the confidence in which one can use a particular prediction is a very important aspect of the application area, it can also facilitate a more optimized data collection approach. More recently, the use of co-kriging methods with external features has been shown to produce a more accurate prediction of isotope

ratios (Truszkowski et al. 2023).

3. **The absence of multimodal feature integration** as existing works do not leverage the vast information present in atmospheric data by systematic feature selection methods. Furthermore, the incorporation of atmospheric variables into the covariance matrix doesn't allow for the same interpretability that decision tree-based methods facilitate.

Our proposed framework uses organic SIRA samples along with atmospheric data to increase feature explainability. We then use feature selection to reduce the dimension of the data and eliminate extraneous information that may deteriorate model performance. For prediction, a combination of prediction and probabilistic models are chosen (1) because of the small sample size, (2) to increase the interpretability of the pipeline when used in real-world applications, and (3) to incorporate atmospheric variables into the prediction task. The spatial pattern of stable isotope predictions is modeled using Gaussian Process (for accuracy) along with a decision tree-boosting algorithm (to fully utilize the feature set). The output of this model is an isoscape with variable values of stable isotope ratios across different geographical locations (West et al. 2009).

This output can then be used to infer the origins of an unseen sample with varying degrees of confidence which has proven to be an effective tool used to assist in the enforcement of trade and sustainability regulations governing commodities like timber, such as identification of sanctioned Russian wood and species listed on the Convention on the International Trade in Endangered Species (Mortier et al. 2024; Grove and Rutherford 2023).

The main contributions of the project are the following:

1. **Reformulating SIRA as an End-to-end ML Pipeline:** We investigate the societally important problem of predicting stable isotope ratios over a large landscape to determine the origin of organic products traded globally and identify illegal harvesting and circumvention of regulations meant to protect resources, such as trees, and the timber and forest products produced from them. To the best of our knowledge, we are the first to create an end-to-end ML prediction pipeline that is usable in practice.
2. **Multifarious data integrating framework that can be applied and interpreted by domain experts:** We propose a novel end-to-end ML Architecture that incorporates atmospheric variables by mapping them to sample locations and then filtering out the extraneous or noisy signals through a feature selection module. The prediction module leverages both the probabilistic modeling of Gaussian process models and the feature extraction power of decision tree-based boosting algorithms.
3. **Experimental validation and benchmarking for future application:** We validate the effectiveness of our model through the means of extensive experimental design. First, we experiment with the optimal number of features from comprehensive atmospheric data. Next, we experiment with different ML techniques which allows us to consider the strengths and weaknesses of each model for real-world use and choose the optimal prediction model. This sets a benchmark and systemic guideline for future researcher on this topic.
4. **Social Impact:** We describe how these methods are already being implemented to have a positive social impact. They are particularly useful in creating isoscape predictions in sub-national areas where it has been impossible to collect physical ground truth samples because of security concerns and sanctions, e.g., in Russia. Reference isoscapes derived from combined ground truth sample data and predicted values for areas lacking samples have been used in 2024 enforcement activities in relation to timber which is at high risk of being Russian harvested in multiple EU Member States.

Related Work

Stable Isotope Ratio Analysis: Stable isotope ratio analysis (SIRA) has been an area of research for several bioorganisms, including but not limited to timber. For example, they have been effective in determining the origin of oceanic creatures like galloprovincialis mussels (del Rio-Lavín et al. 2022), fish (Cusa et al. 2022) or dairy products (O'Sullivan, Schmidt, and Monahan 2022). While these works have demonstrated the effectiveness of traditional statistical modeling for stable isotope ratio modeling, they generally focus on smaller geographical areas and are not limited by the data paucity issues that conventional ML models often face. When it comes to SIRA on timber, prior works have shown promising results with a simplistic modeling approach that does not generalize to a larger geographical area (Watkinson et al. 2020, 2022). There still remains scope to incorporate external factors that inform the spatial-organic process of these organisms as Truszkowski et al. (2023) and Mortier et al. (2024) have shown. Motivated by these findings, we propose a holistic feature integration pipeline that aims to further capture the non-linear, hierarchical nature of these external factors.

Spatial Regression: Spatial predictive modeling, despite being a widely researched topic, has potential to utilize the rapid progress in multimodal data collection framework. While kriging has been used for ubiquitous tasks like sound-level mapping (Aumond et al. 2018), spatial interpolation of horizon depth (Knotters, Brus, and Voshaar 1995), spatial modeling of water quality index (Khan et al. 2023), and other tasks, Gaussian process regression (GPR) has proven to be effective for spatial prediction of drought (Elbeltagi et al. 2023), standardized precipitation index, and prediction of COVID-19 spread (Velásquez and Lara 2020). Furthermore, techniques like inverse distance weighting (IDW) has also been shown to be effective for tasks such as evaluating groundwater quality (Singh and Verma 2019), reference evaporation (Hodam et al. 2017), rainfall distribution estimation (Chen and Liu 2012). However, the inherent assumption of these models about the spatial homogeneity of the target variable may not extend to stable isotopes as external factors like temperature, soil, and altitude may affect the values. Hence, we leverage the utility of spatial prediction models in capturing the geographical variance along with

the added feature of incorporating uncertainty with a more streamlined incorporation of additional features.

ML in application: The application of conventional ML in the application of fields such as biological sciences, epidemiology, and meteorology is ubiquitous (Velásquez and Lara 2020; Rustam et al. 2020; Maurya et al. 2023; Elbeltagi et al. 2023; Balaji 2021). One key takeaway from this vast array of works, has been the effectiveness of ML models to automate the process of feature selection as demonstrated by Sarkar, Alhamadani, and Lu (2022), Maurya et al. (2023) and other works. We follow a similar approach of selecting features from hundreds of variables before feeding them into the predictive model. One key challenge in SIRA is the lack of samples with reliable values for each stable isotope, making large-scale ML and DL models unsuitable. Therefore, decision tree-based models and boosting algorithms are ideal candidates to leverage the selected atmospheric variables. To this effect Pan et al. (2023), showed how relevant features can be learned by combining multiple weak decision tree-based learners. When presented with a structured selection of relevant features decision tree-based models have faired well even against larger black-box DL models, leading to a more sustainable application of AI techniques (Ferro et al. 2023). However, due to the real world applicability of SIR prediction, there is an added importance to quantifying uncertainty as demonstrated by probabilistic modeling techniques like the Gaussian Process or Kriging. However, it can be combined with decision tree-boosting algorithms to leverage both the probabilistic nature of spatial correlation and the non-linear correlation of structured high-dimensional feature vectors (Sigrist 2022).

Preliminaries

Stable Isotope Ratios: A stable isotope ratio is the ratio of two stable isotopes of a single element (Coplen, Kendall, and Hopple 1983). The measurement is usually expressed through delta notation that signifies percentage per milliliter. Oxygen, carbon, hydrogen, nitrogen, and sulfur are some of the most common elements whose stable isotope ratios are measured and used for various scientific analyses. For each element, we have a different stable isotope that we are interested in. In general, it can be defined as follows:

$$\delta^{High}\mathrm{E} = \left(\frac{\left({}^{high}\mathrm{E}/{}^{natural}\mathrm{E}\right)_{\text{sample}}}{\left({}^{high}\mathrm{E}/{}^{natural}\mathrm{E}\right)_{\text{standard}}} - 1 \right) \times 1000$$

where E is the element. The following are the four stable isotope ratios used in this paper:

- $\delta^{13}C$: a measure of the ratio of two stable isotopes of Carbon (^{13}C and ^{12}C).
- $\delta^{18}O$: a measure of the ratio of two stable isotopes of Oxygen (^{18}O and ^{16}O).
- $\delta^{2}H$: a measure of the ratio of two stable isotopes of Hydrogen (^{2}H and ^{1}H).
- $\delta^{34}S$: a measure of the ratio of two stable isotopes of Sulfur (^{34}S and ^{32}S).

Problem Statement: Given a set of n samples with location X = X_1, X_2, X_3...X_n with corresponding atmospheric variable A(X)=$A(X_1)$,$A(X_2)$,$A(X_3)$,...,$A(X_n)$ we aim to learn a function $Y = F(A(X), X)$ where Y corresponds to a single stable isotope ratio value. Hence, for each isotope ratio we aim to train an individual model and predict stable isotope ratio values over an isoscape.

The Proposed Framework

The GB-SIRA pipeline for stable isotope ratio prediction of timber is designed as an end-to-end ML Framework with each stage performing a task, the output of which is then used as input for the next stage (Fig. 1). The first stage is the atmospheric feature construction module. This module uses 25 different datasets containing the last 20 years of atmospheric data for locations across the globe and maps it to the locations associated with the timber samples in the dataset. The next module of feature selection takes the feature vectors of atmospheric variables and selects top-k pertinent features which is used to train the optimal model for prediction of isotope ratios.

Feature Construction: Stable isotope ratios are largely dependent on various environmental factors including but not limited to precipitation, water vapor pressure, reflected short wave radiation(Vystavna, Matiatos, and Wassenaar 2021; Elbeltagi et al. 2023), and the ecological process that facilitates the enrichment, or lack thereof, for stable isotopes as observed in oak trees is a lengthy process that can be captured by climatic patterns over a long time. Hence, the collected comprehensive data about these atmospheric properties(NASA 2024; Huffman et al. 2020; Bowen and Revenaugh 2003) is further aggregated by month for 20 years to capture the overall climatic condition and the periodic pattern. The dataset for oak trees includes stable isotope ratio values along with the latitude and longitude where the oak wood sample was collected. Coordinates common in both *Quercuss spp.* and atmospheric can be directly mapped. However, for coordinates not present in the atmospheric dataset, consistent with Tobler's first law of geography.(Miller 2004), we use IDW interpolation with a threshold geodesic distance.

Feature Selection Module: After aggregating and mapping these variables to every location and eliminating variables with more than 50 percent null values we are left with a feature vector longer than 200. Since the dataset has 487 samples, we focus on reducing the dimension using the F-regression technique. The hyperparameter in this case is the choice of k which indicates several features. However, regardless of the value of k, the main objective of this technique is choosing the most correlated features. Since, in our work, we are interested in 4 different isotope ratios, our target variables are δ^{13}C, δ^{2}H, δ^{2}H, and δ^{18}O. For each target and independent variable, the F-value is calculated which is then used to rank the features. Thus for each stable isotope ratio, we choose the top-k atmospheric variables as feature vectors of length k. In our experiments, we found $k = 21$ to be optimal for δ^{13}C, $k = 19$ to be optimal for δ^{2}H, k= 25 to be optimal for δ^{18}O, and $k = 30$ to be optimal for δ^{2}H.

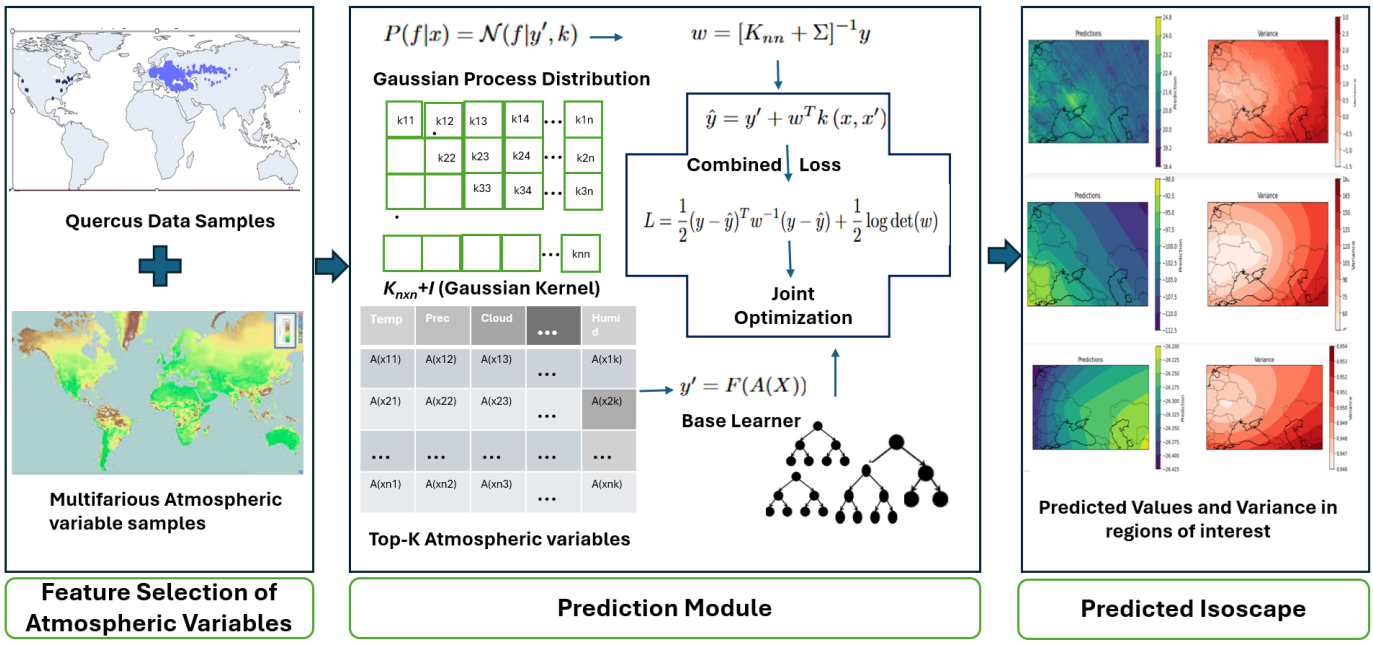

Figure 1: GB-SIRA framework

Prediction Module: In this section we are tasked with leveraging two types of data: 1) spatial data in the form of latitude and longitude values for the oak wood samples, and 2) atmospheric feature data as extracted from the feature selection module. Gaussian Process Models are often used for modeling spatial data which is also known as the universal kriging method. However, for capturing the complex non-linear relationship with the atmospheric variables, mixed effect models like decision tree-based boosting algorithms have shown to be effective. In this section, we will describe the two methods and discuss how the combined method of Gaussian process and mixed effect modeling approach as described in(Sigrist 2022) has been implemented for our prediction module. Given we have a set of locations X=x_1,x_2...,x_n where the corresponding atmospheric feature vector is $A(X)$=$a(x_1)$,$a(x_1)$,...,$a(x_n)$ and $a(X_i)$ is the atmospheric feature at location x_i, we want to learn two separate learners, one for the atmospheric data and one for the spatial data. First, let's talk about the mixed-effect learner.

Learning Base Learners: To model the non-linear relationship between atmospheric variables and stable isotopes, we use Mixed-Effect Modeling consisting 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: GPR is particularly popular for scientific applications because of its probabilistic modeling approach that allows for uncertainty estimation. The main assumption is the data samples as observed in the dataset are described by a set of functions instead of a single function with set parameters. Hence, it is an inherently non-parametric model. Now, the probability density function that forms the basis of this model is defined as $P(f|x) = \mathcal{N}(f|\mu, k)$ here $\mu = [m(x1), m(x2)...m(xn)]$ is the mean estimate and K is the covariance matrix where $K_{ij} = k(x_i, x_j)$ where k is kernel covariance function. Traditionally, it 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 as an input $X = x_1, x_2, x_3...x_n$ where x_i denotes GPS location of ith sample in the training set. The target variable y is a single isotope ratio. The GPR fits the following kernel coefficient W:

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

where Σ is the error term that corresponds to the uncertainty estimation and K_{nn} is the covariance matrix of dimension (n x n) where n is the number of samples in the training set.

Now, for a new set of unseen samples x, the prediction will be calculated as

$$\hat{y} = y' + w^T k(x, x') \tag{3}$$

In our case, K is the kernel function of our choosing. We are using a combination of radial basis function, periodic, and rational quadratic function as the covariance function as

shown below:

$$k\left(x,x'\right)=\begin{bmatrix}\exp\left(-\frac{\left(x-x^1\right)^2}{2\ell^2}\right)+\left(1+\frac{\left(x-x^1\right)^2}{2\alpha\ell^2}\right)^{-\alpha}\\ +\exp\left(-\frac{2\sin^2\left(\pi\left|x-x^1\right|/p\right)}{\ell^2}\right),\\ \exp\left(-\frac{\left(x-x^2\right)^2}{2\ell^2}\right)+\left(1+\frac{\left(x-x^2\right)^2}{2\alpha\ell^2}\right)^{-\alpha}\\ +\exp\left(-\frac{2\sin^2\left(\pi\left|x-x^2\right|/p\right)}{\ell^2}\right),\\ \vdots\\ \exp\left(-\frac{(x-x^n)^2}{2\ell^2}\right)+\left(1+\frac{(x-x^n)^2}{2\alpha\ell^2}\right)^{-\alpha}\\ +\exp\left(-\frac{2\sin^2(\pi|x-x^n|/p)}{\ell^2}\right)\end{bmatrix} \tag{4}$$

Joint Optimization: As described in (Sigrist 2022) we need to combine the two learners through a joint optimization goal. Here the loss function is the negative log-likelihood function as described below:

$$L=\frac{1}{2}(y-\hat{y})^T w^{-1}(y-\hat{y})+\frac{1}{2}\log\det(w) \tag{5}$$

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

$$(\hat{F}(\cdot),\hat{\mathbf{w}})=\underset{(F(\cdot),\mathbf{w})\in(\mathcal{H},\mathbf{w})}{\operatorname{argmin}} R(F(\cdot),\mathbf{w}) \tag{6}$$

where $R(F(\cdot),\mathbf{w})$ is a risk functional defined as:

$$R(F(\cdot),\mathbf{w}):\quad (F(\cdot),\mathbf{w})\mapsto L \tag{7}$$

where 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 $\mathbf{w}_i$ as the following:

$$\mathbf{w}_i=\operatorname{argmin}_{\mathbf{w}\in\mathbf{w}} L\left(y,F_{i-1},\mathbf{w}\right) \tag{8}$$

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

$$F_i=F_{i-1}-\frac{f'\left(F_{i-1}\right)}{f''\left(F_{i-1}\right)} \tag{9}$$

where f' and f'' denotes first and second order derivatives.

Experiment and Results

Data Description and Preprocessing

***Quercus* spp. dataset:** We use data from 487 trees of the genus *Quercus* with samples 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 (Watkinson et al. 2020). Each entry contained stable isotope ratio measurements of oxygen, hydrogen, sulfur, and carbon and the samples GPS coordinates.

Atmospheric dataset and cleaning: Our atmospheric data includes isotopic composition of precipitation, water vapour, shortwave radiation, temperature, and many more other factors (Bowen and Revenaugh 2003; Huffman et al. 2020; NASA 2024). This dataset consists of 25 atmospheric variables for 20 years. However, while mapping the atmospheric variable to the sampled location of the *Quercus* spp. dataset, we found a few of the variables to have more than 50 percent values null because of the sparse nature of the data. That left 19 atmospheric variables each having 12 months of data for every year. We use the dataset of collected samples to map atmospheric variables to the location of a given data point, allowing us to create a comprehensive feature set of atmospheric variables for each sample in the dataset along with their location.

Experimental Research Questions

The research questions related to the experimental evidence regarding the validity of the design and methodology of the architecture to understand the effectiveness of the proposed GB-SIRA framework, are the following:

- **RQ1:** How does the proposed framework compare to the existing works in stable isotope ratio prediction?
- **RQ2:** Does combining tree boosting with Gaussian Process through Gaussian Process Boosting benefit the outcome compared with using either of the techniques individually?
- **RQ3:** Is the choice of feature selection and incorporation of atmospheric variables justified by more accurate prediction?

Experimental Setup and Evaluation

For all isotope ratios we split the dataset into an 80:20 training/test split. We perform K-fold cross-validation with $k = 5$ and then 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 and max depth to be 3.

The model is evaluated using standard metrics for regression tasks, R^2 value and RMSE score. The definitions of both are given below:

R^2 **value:** the R-squared 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: RMSE helps understand how close the prediction is to the ground truth across the test samples.

RQ1: Baseline Experiments

To answer the first research question, we compare our framework to existing work in stable isotope ratio prediction and compare our results based on the chosen metrics. For the works that did not use the same set of sample location, we employ their methodology on our data for comparison

- Watkinson et al. (2020): This work proposes the use of ordinary kriging for the spatial interpolation of the isotope ratio values.

Baseline Comparison	Atmospheric Variables	R^2				RMSE			
		δ^{18}O	δ^{13}C	δ^{2}H	$\delta^{34}S$	δ^{18}O	δ^{13}C	δ^{2}H	$\delta^{34}S$
Watkinson et al. (2020)		0.470	0.320	0.700	**0.690**	-	-	-	-
Watkinson et al. (2022)		0.856	0.301	0.730	0.601	0.673	0.779	6.112	1.233
Truszkowski et al. (2023)	X	0.869	**0.331**	0.790	0.667	0.631	0.757	6.279	1.070
RF	X	0.894	0.284	0.779	0.553	0.660	0.781	6.380	1.310
SVR	X	0.754	0.313	0.750	0.682	0.940	0.773	6.410	1.040
GB-SIRA	X	**0.902**	0.322	**0.841**	0.689	**0.619**	**0.768**	**5.911**	**1.011**

Table 1: Baseline Comparison Results

- Watkinson et al. (2022): This work uses a GPR to predict the isotope ratio values in *Quercus* spp.
- Truszkowski et al. (2023): 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.
- Random Forest (RF): Among decision tree-based methods RF has traditionally been shown to be effective for modeling structured data where the linear assumption of the model may not hold.
- 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.

Results: The baseline experiments as detailed in Table 1 show the performance of our proposed model GB-SIRA compared to others. From the results, it is clear that in terms of R^2 value, GB-SIRA outperforms the state-of-the-art for 3 of the 4 isotope ratio prediction tasks. However, even for δ^2H the difference with the best-performing model is negligible. It is important to note that for Watkinson et al. 2020, RMSE was not reported. In terms of RMSE, we also observe the lowest RMSE value for GB-SIRA for all stable isotope ratios. The performance of the GB-SIRA overall compares favorably against existing works on the same task. This shows the effectiveness of the GB-SIRA framework.

RQ2: Does combining tree-boosting with GPR benefit the model performance?

To answer this research question we tested each individual element after the feature selection module to compare against the final model. The result of this experimental setting is described in Table 2. 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. When compared to the performance of the proposed framework, these two models do not show any improvement in the R_2 despite showing slight improvement on RMSE for lighGBM. However, given the significance of uncertainty estimation for real-world applications, it is clear that combining the two methods is the right strategy.

Ablation Study	R^2			
	δ^{18}O	δ^{13}C	δ^{2}H	$\delta^{34}S$
Boosting	0.885	0.329	0.819	0.675
GPR	0.856	0.301	0.730	0.601
GB-SIRA	**0.902**	**0.322**	**0.841**	**0.689**
	RMSE			
Boosting	0.601	0.780	5.820	1.070
GPR	0.673	0.779	6.112	1.233
GB-SIRA	**0.619**	**0.768**	**5.911**	**0.987**

Table 2: Ablation Study

RQ3: Is the choice of feature selection justified?

To answer this we look at the two best-performing models from Table 1. Since the co-kriging method also takes advantage of the atmospheric variable, we create an experimental setting that facilitates both the inclusion and exclusion of the feature selection module. The result as described in Table 3 shows that the inclusion of the feature selection step before modeling produces a better prediction for both models.

Case Study and Social Impact

The real-world implication of having GB-SIRA pipeline 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.

These isoscape predictions have been made around the Russia/Ukraine border region where it has been impossible to collect physical ground truth samples because of security concerns and sanctions as seen in Fig 2. These isoscapes have then been leveraged in the past year for the enforcement of timber trade, which is at high risk of being Russian harvested in multiple EU Member States (Tokar 2024; Nazaryan 2024).

Our work contributes to social and environmental impact initiatives by contributing much-needed advances to shed

	R^2				RMSE			
	δ^{18}O	δ^{13}C	δ^{2}H	$\delta^{34}S$	δ^{18}O	δ^{13}C	δ^{2}H	$\delta^{34}S$
Co-Kriging without Feature Selection	0.850	0.271	0.766	0.631	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
Co-Kriging with Feature Selection	0.869	**0.331**	0.790	0.667	0.631	**0.757**	6.279	1.070
GB-SIRA with Feature Selection	**0.902**	0.322	**0.841**	**0.689**	**0.619**	0.768	**5.911**	**0.987**

Table 3: Experiments with Feature Selection

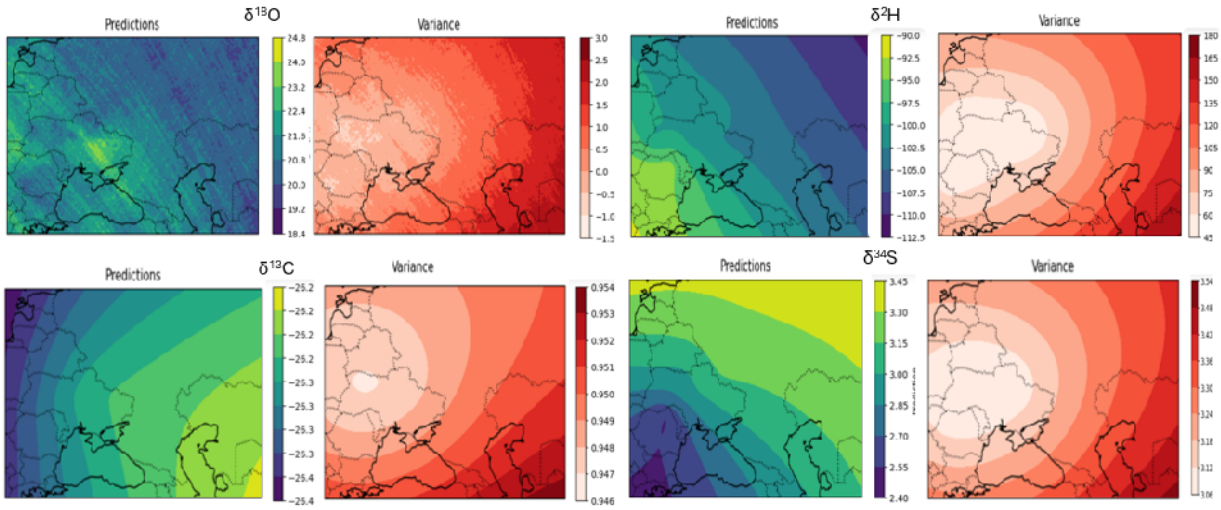

Figure 2: Case Study: Isoscapes for Eastern European Region-Ukraine/Russia Border for timber

light on complex global supply chains and tools that can be used to ensure the products we consume do not contribute to social or environmental harm. These methods can be used to verify that wood and forest products are labeled correctly and come from legal and sustainable sources. It can also be used to identify timber harvested illegally in protected areas, in high conservation value forests, or in the lands of Indigenous peoples and vulnerable communities that might be mislabeled to obfuscate the true harvest origin.

Conclusion

We presented a holistic multimodal ML framework for stable isotope ratio prediction that also estimates uncertainty for each prediction which allows domain experts to use it in a real-world setting. We demonstrate the quantitative effectiveness of combining atmospheric variables through nonlinear hierarchical modeling facilitated by decision tree boosting and spatial probabilistic modeling of the Gaussian Process, by comparing our results with other related work and ablation studies. Furthermore, the case study showcases that the application of this work is already having a social impact. Isoscapes produced by GB-SIRA are being used to verify claims made about the origin of organic products and helping industry and government officials identify instances of false or fraudulent location-of-harvest claims.

Future Research: We are in the process of extending our work to identify origin of grains in regions where collection of data is hard to come by. This work will focus on going beyond stable isotope ratio prediction by developing a pipeline for the inverse problem of location determination.

Further Application: 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 is particularly relevant given the EU's regulation on deforestation-free products and forest-risk commodities (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 (European Commission 2023).

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 (McLymore 2024; Masters 2023; PBS News 2023; Cusa et al. 2022). 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.

References

Aumond, P.; Can, A.; Mallet, V.; De Coensel, B.; Ribeiro, C.; Botteldooren, D.; and Lavandier, C. 2018. Kriging-based spatial interpolation from measurements for sound level mapping in urban areas. *The journal of the acoustical society of America*, 143(5): 2847–2857.

Balaji, V. 2021. Climbing down Charney's ladder: machine learning and the post-Dennard era of computational climate science. *Philosophical Transactions of the Royal Society A*, 379(2194): 20200085.

Barrie, A.; and Prosser, S. 1996. Automated analysis of light-element stable isotopes by isotope ratio mass spectrometry. *Mass spectrometry of soils. New York, Marcel Dekker*, 1–46.

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

Chen, F.-W.; and Liu, C.-W. 2012. Estimation of the spatial rainfall distribution using inverse distance weighting (IDW) in the middle of Taiwan. *Paddy and Water Environment*, 10: 209–222.

Coplen, T. B.; Kendall, C.; and Hopple, J. 1983. Comparison of stable isotope reference samples. *Nature*, 302(5905): 236–238.

Cusa, M.; St John Glew, K.; Trueman, C.; Mariani, S.; Buckley, L.; Neat, F.; and Longo, C. 2022. A future for seafood point-of-origin testing using DNA and stable isotope signatures. *Reviews in Fish Biology and Fisheries*, 32(2): 597–621.

del Rio-Lavín, A.; Weber, J.; Molkentin, J.; Jiménez, E.; Artetxe-Arrate, I.; and Pardo, M. Á. 2022. Stable isotope and trace element analysis for tracing the geographical origin of the Mediterranean mussel (Mytilus galloprovincialis) in food authentication. *Food Control*, 139: 109069.

Elbeltagi, A.; Pande, C. B.; Kumar, M.; Tolche, A. D.; Singh, S. K.; Kumar, A.; and Vishwakarma, D. K. 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): 43183–43202.

European Commission. 2023. Regulation on Deforestation-free products.

Ferro, M.; Silva, G. D.; de Paula, F. B.; Vieira, V.; and Schulze, B. 2023. Towards a sustainable artificial intelligence: A case study of energy efficiency in decision tree algorithms. *Concurrency and Computation: Practice and Experience*, 35(17): e6815.

Grove, M.; and Rutherford, C. 2023. CITES and Timber: a guide to CITES-listed tree species. Technical report, CITES.

Hodam, S.; Sarkar, S.; Marak, A. G.; Bandyopadhyay, A.; and Bhadra, A. 2017. Spatial interpolation of reference evapotranspiration in India: Comparison of IDW and Kriging methods. *Journal of the Institution of Engineers (india): Series A*, 98: 511–524.

Huffman, G.; Behrangi, A.; Bolvin, D.; and Nelkin, E. 2020. GPCP version 3.1 satellite-gauge (SG) combined precipitation data set. *NASA GES DISC: Greenbelt, MD, USA*.

Khan, M.; Almazah, M. M.; EIlahi, A.; Niaz, R.; Al-Rezami, A.; and Zaman, B. 2023. Spatial interpolation of water quality index based on Ordinary kriging and Universal kriging. *Geomatics, Natural Hazards and Risk*, 14(1): 2190853.

Knotters, M.; Brus, D.; and Voshaar, J. O. 1995. A comparison of kriging, co-kriging and kriging combined with regression for spatial interpolation of horizon depth with censored observations. *Geoderma*, 67(3-4): 227–246.

Masters, K. 2023. US Customs finds garments made with banned Chinese cotton. *Reuters*.

Maurya, N. S.; Kushwah, S.; Kushwaha, S.; Chawade, A.; and Mani, A. 2023. Prognostic model development for classification of colorectal adenocarcinoma by using machine learning model based on feature selection technique boruta. *Scientific reports*, 13(1): 6413.

McLymore, A. 2024. Banned Chinese cotton found in 19% of US and global retailers' merchandise, study shows. *Reuters*.

Meier-Augenstein, W.; Kemp, H. F.; Schenk, E. R.; and Almirall, J. R. 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): 545–552.

Miller, H. J. 2004. Tobler's first law and spatial analysis. *Annals of the association of American geographers*, 94(2): 284–289.

Mortier, T.; Truszkowski, J.; Norman, M.; Boner, M.; Buliga, B.; Chater, C.; Jennings, H.; Saunders, J.; Sibley, R.; Antonelli, A.; et al. 2024. A framework for tracing timber following the Ukraine invasion. *Nature Plants*, 10(3): 390–401.

NASA. 2024. Reflected Shortwave Radiation neo.gsfc.nasa.gov. Accessed 15-08-2024.

Nazaryan, A. 2024. New Method That Pinpoints Wood's Origin May Curb Illegal Timber. *The New York Times*.

O'Sullivan, R.; Schmidt, O.; and Monahan, F. J. 2022. Stable isotope ratio analysis for the authentication of milk and dairy ingredients: A review. *International Dairy Journal*, 126: 105268.

Pan, R.; Liu, T.; Huang, W.; Wang, Y.; Yang, D.; and Chen, J. 2023. State of health estimation for lithium-ion batteries based on two-stage features extraction and gradient boosting decision tree. *Energy*, 285: 129460.

PBS News. 2023. Investigation reveals Chinese seafood caught and processed using forced labor sold in U.S. Section: World.

Rustam, F.; Reshi, A. A.; Mehmood, A.; Ullah, S.; On, B.-W.; Aslam, W.; and Choi, G. S. 2020. COVID-19 future forecasting using supervised machine learning models. *IEEE access*, 8: 101489–101499.

Sarkar, S.; Alhamadani, A.; and Lu, C.-T. 2022. Explainable Prediction of the Severity of COVID-19 Outbreak for US

Counties. In *2022 IEEE International Conference on Big Data (Big Data)*, 5338–5345. IEEE.

Siegwolf, R. T. W.; Brooks, J. R.; Roden, J.; and Saurer, M., eds. 2022. *Stable Isotopes in Tree Rings: Inferring Physiological, Climatic and Environmental Responses*, volume 8 of *Tree Physiology*. Cham: Springer International Publishing. ISBN 978-3-030-92697-7 978-3-030-92698-4.

Sigrist, F. 2022. Gaussian process boosting. *Journal of Machine Learning Research*, 23(232): 1–46.

Singh, P.; and Verma, P. 2019. A comparative study of spatial interpolation technique (IDW and Kriging) for determining groundwater quality. *GIS and geostatistical techniques for groundwater science*, 43–56.

Tokar, D. 2024. Keeping Sanctioned Russian Timber Out of the EU Is Tricky. This Nonprofit Has a Solution. *Wall Street Journal*.

Truszkowski, J. M.; Maor, R.; Yousuf, R. B.; Biswas, S.; Chater, C.; Gasson, P.; McQueen, S.; Norman, M.; Saunders, J.; Simeone, J.; et al. 2023. A probabilistic approach to estimating timber harvest location. *EcoEvoRxiv*.

Velásquez, R. M. A.; and Lara, J. V. M. 2020. Forecast and evaluation of COVID-19 spreading in USA with reduced-space Gaussian process regression. *Chaos, Solitons & Fractals*, 136: 109924.

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

Wang, J.; Chen, T.; Zhang, W.; Zhao, Y.; Yang, S.; and Chen, A. 2020. Tracing the geographical origin of rice by stable isotopic analyses combined with chemometrics. *Food chemistry*, 313: 126093.

Wang, L.; Jin, Y.; Weiss, D. J.; Schleicher, N. J.; Wilcke, W.; Wu, L.; Guo, Q.; Chen, J.; O'Connor, D.; and Hou, D. 2021. Possible application of stable isotope compositions for the identification of metal sources in soil. *Journal of Hazardous Materials*, 407: 124812.

Watkinson, C. J.; Gasson, P.; Rees, G. O.; and Boner, M. 2020. The development and use of isoscapes to determine the geographical origin of Quercus spp. in the United States. *Forests*, 11(8): 862.

Watkinson, C. J.; Rees, G. O.; Hofem, S.; Michely, L.; Gasson, P.; and Boner, M. 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: 645222.

West, J. B.; Bowen, G. J.; Dawson, T. E.; and Tu, K. P. 2009. *Isoscapes: Understanding movement, pattern, and process on Earth through isotope mapping*. Springer Science & Business Media. ISBN 978-90-481-3354-3. Google-Books-ID: XYqq0zT7em8C.

Yang, C.; Hutcheon, I.; and Krouse, H. 2001. Fluid inclusion and stable isotopic studies of thermochemical sulphate reduction from Burnt Timber and Crossfield East gas fields in Alberta, Canada. *Bulletin of Canadian Petroleum Geology*, 49(1): 149–164.

Reproducibility Checklist:

This paper:

- Includes a conceptual outline and/or pseudocode description of AI methods introduced: Yes
- Clearly delineates statements that are opinions, hypothesis, and speculation from objective facts and results (yes)
- Provides well marked pedagogical references for less-familiare readers to gain background necessary to replicate the paper (yes)

Does this paper make theoretical contributions? (no)

Does this paper rely on one or more datasets? (yes) If yes, please complete the list below.

- A motivation is given for why the experiments are conducted on the selected datasets (yes)
- All novel datasets introduced in this paper are included in a data appendix. (partial)
- All novel datasets introduced in this paper will be made publicly available upon publication of the paper with a license that allows free usage for research purposes. (yes)
- All datasets drawn from the existing literature (potentially including authors' own previously published work) are accompanied by appropriate citations. (yes)
- All datasets drawn from the existing literature (potentially including authors' own previously published work) are publicly available. (partial) datasets that are not publicly available are described in detail, with explanation why publicly available alternatives are not scientifically satisficing. (yes)

Does this paper include computational experiments? (yes) If yes, please complete the list below.

- Any code required for pre-processing data is included in the appendix. (partial).
- All source code required for conducting and analyzing the experiments is included in a code appendix. (partial)
- All source code required for conducting and analyzing the experiments will be made publicly available upon publication of the paper with a license that allows free usage for research purposes. (yes)
- All source code implementing new methods have comments detailing the implementation, with references to the paper where each step comes from (yes)
- If an algorithm depends on randomness, then the method used for setting seeds is described in a way sufficient to allow replication of results. (yes)
- This paper specifies the computing infrastructure used for running experiments (hardware and software), including GPU/CPU models; amount of memory; operating system; names and versions of relevant software libraries and frameworks. (partial)
- This paper formally describes evaluation metrics used and explains the motivation for choosing these metrics. (yes)
- This paper states the number of algorithm runs used to compute each reported result. (yes)

- Analysis of experiments goes beyond single-dimensional summaries of performance (e.g., average; median) to include measures of variation, confidence, or other distributional information. (no)
- The significance of any improvement or decrease in performance is judged using appropriate statistical tests (e.g., Wilcoxon signed-rank). (yes)
- This paper lists all final (hyper-)parameters used for each model/algorithm in the paper's experiments. (partial)
- This paper states the number and range of values tried per (hyper-) parameter during development of the paper, along with the criterion used for selecting the final parameter setting. (partial)

Appendix:

Programming Language Used: Python.

Libraries used:SkLearn, gpboost, shapely.geometry, pykrige, pandas, numpy,lightgbm

Hyperparamters: For

- SVR: Kernel -SVR, degree = 3
- RF: $random_state$ = 0
- GPBoost: kernel- custom(as described in main text), max_depth = 3, $learning_rate$ = 0.03
- Gaussian Process: kernel = custom
- LightgbM: max_depth = 5
- ordinary kriging: maxlag = 2, n_lags=15

Code and Data: Entire codebase and required data is shared in supplementary files