Optimizing Product Provenance Verification Using Data Valuation Methods

Raquib Bin Yousuf[1], Hoang Anh Just[1], Shengzhe Xu[1], Brian Mayer[1], Victor Deklerck[2], Jakub Truszkowski[3,4], John C. Simeone[5], Jade Saunders[4], Chang-Tien Lu[1], Ruoxi Jia[1], Naren Ramakrishnan[1]

[1]Virginia Tech, VA, USA
[2]Meise Botanic Garden, Meise, Belgium
[3]Chalmers University of Technology, Gothenburg, Sweden
[4]World Forest ID, Washington, DC, USA
[5]Simeone Consulting, LLC, Littleton, NH, USA
raquib@vt.edu, naren@cs.vt.edu

Abstract

Determining and verifying product provenance remains a critical challenge in global supply chains, particularly as geopolitical conflicts and shifting borders create new incentives for misrepresentation of commodities, such as hiding the origin of illegally harvested timber or stolen agricultural products. Stable Isotope Ratio Analysis (SIRA), combined with Gaussian process regression-based isoscapes, has emerged as a powerful tool for geographic origin verification. While these models are now actively deployed in operational settings supporting regulators, certification bodies, and companies, they remain constrained by data scarcity and suboptimal dataset selection. In this work, we introduce a novel deployed data valuation framework designed to enhance the selection and utilization of training data for machine learning models applied in SIRA. By quantifying the marginal utility of individual samples using Shapley values, our method guides strategic, cost-effective, and robust sampling campaigns within active monitoring programs. By prioritizing high-informative samples, our approach improves model robustness and predictive accuracy across diverse datasets and geographies. Our framework has been implemented and validated in a live provenance verification system currently used by enforcement agencies, demonstrating tangible, real-world impact. Through extensive experiments and deployment in a live provenance verification system, we show that this system significantly enhances provenance verification, mitigates fraudulent trade practices, and strengthens regulatory enforcement of global supply chains.

Extended version — https://arxiv.org/abs/2502.15177

1 Introduction

Global natural resource supply chains are opaque, especially since natural resources are often transformed from raw materials (e.g., timber) into finished consumer-facing products (e.g., furniture). These complex supply chains often involve multiple countries, with intermediate outputs being traded internationally and being used as inputs into further manufacturing processes. In addition to business-commerce decisions driving supply chain sourcing, the economics of natural resource trade are often closely linked with geopolitics. Verifying the true origin of products remains difficult and high-stakes, as geopolitical incentives and "don't ask, don't tell" sourcing cultures encourages misrepresentation of commodities, particularly for illegally harvested timber or stolen agricultural products.

The Scale of the Problem: Illegal Timber as a Case Study. Illegal logging, in particular, is the most profitable natural resource crime, valued at US$52 billion to US$157 billion per year (May 2017). Illegally obtained timber accounts for 10–30% of the total global trade in timber products, and in regions such as Southeast Asia, Central Africa, and South America, it is estimated that 50–90% of timber is harvested illegally (May 2017). These figures underscore the urgent need for deployed, verifiable, and scientifically defensible provenance systems.

Forensic and Analytical Tools for Provenance Verification. Product identification and provenance verification of traded natural resources have emerged as promising research areas, with various combinations of methods used depending on the resource sector and the granularity required for species identification and origin determination; for example, species and geographic harvest provenance for wood and forest products often requires multiple forensic tools and testing methods (Grant and Chen 2021; Schmitz et al. 2020; Dormontt et al. 2015). For geographic origin verification in particular, Stable Isotope Ratio Analysis (SIRA), combined with Gaussian process regression-based isoscapes, has proven highly effective (Truszkowski et al. 2025; Mortier et al. 2024). SIRA leverages the natural variation in stable isotope ratios—measured via mass spectrometry—to determine the enrichment of non-radioactive elemental isotopes in a sample (Barrie and Prosser 1996), with enrichment patterns driven by environmental, atmospheric, soil, metabolic, and species-specific factors (Siegwolf et al. 2022; Wang et al. 2021; Vystavna, Matiatos, and Wassenaar 2021). It has been successfully used to trace the origin of timber, seafood, agricultural products, and fiber such as cotton (Truszkowski et al. 2025; Mortier et al. 2024; Watkinson et al. 2022; Cusa et al. 2022; Wang et al. 2020; Meier-Augenstein et al. 2014). However, operational deployment

of these models faces challenges of data scarcity, suboptimal reference selection, and high sampling costs, motivating the need for systematic data valuation and prioritization. This in turn highlights the importance of optimizing and 'valuing' reference sample collection efforts (Gasson et al. 2021).

Operationalizing Provenance Verification: The WFID Platform. Efforts such as World Forest ID (WFID) (worldforestid.org) have operationalized geographic origin models to address this problem, particularly in the timber trade. WFID identifies high-risk species and geographies (Norman 2023), collects chain-of-custody reference samples, and uses laboratory analysis (Fera Science 2025) to generate the chemical data needed for stable isotope modeling. These models are not only research prototypes—they are currently deployed and used in live compliance systems. The World Forest ID Evaluation Platform (World Forest ID 2024b) allows regulators, certification bodies, and companies to evaluate sourcing claims in real time.

Here, we introduce our deployed data valuation framework that enhances the selection and utilization of training data for machine learning models applied to SIRA. Field sample collections, central to any scientific traceability method, are logistically difficult and expensive due to remote locations, specialized equipment, and labor-intensive workflows. Our deployed system operationalizes Shapley-based data valuation to quantify the marginal utility of individual samples, enabling strategic, cost-effective, and robust sampling campaigns that directly improve model accuracy in active enforcement workflows.

Contributions:

1. Application of machine learning data valuation to an operational provenance verification context. By prioritizing highly informative samples, our approach improves model robustness and predictive accuracy across diverse datasets and geographies.
2. Deployment in live enforcement systems. We have deployed our optimized sampling approach in a provenance verification system used by European enforcement agencies to curb the trade of sanctioned Russian timber by proving that alternate claimed origins are non-viable. See coverage of our work in the *New York Times* (Nazaryan 2024)). Due to confidentiality reasons, we use a global dataset of Oak (*Quercus spp.*) reference samples to illustrate our methodology.
3. Extensive empirical validation and field integration. We validate our framework with global Oak (Quercus spp.) reference datasets, demonstrating its potential to enhance data valuation, optimize model configuration, and strengthen the regulatory enforcement of global supply chains.

2 Related Work

SIRA has been widely employed as a geographical discriminator for various plant and animal-based products in global supply chains, such as garlic (Pianezze et al. 2019), Chinese tea (Liu et al. 2020), olive oil (Bontempo et al. 2019), cheese (Camin et al. 2004), and timber (Mortier et al. 2024; Truszkowski et al. 2025). By bringing data valuation methods to bear upon SIRA pipelines we aim to improve the verification of product provenance.

Prior work in data valuation is typically seen in the context of explainable machine learning and enhancing model performance (Wu, Zhu, and Li 2024; Covert et al. 2024). Existing methods primarily rely on leave-one-out retraining and influence functions (Koh and Liang 2017), Shapley values (Jia et al. 2019; Ghorbani and Zou 2019; Wang and Jia 2023), Least Cores (Yan and Procaccia 2021), the Banzhaf value (Wang and Jia 2022), Beta Shapley (Kwon and Zou 2021), and reinforcement learning (Yoon, Arik, and Pfister 2020). Furthermore, data valuation has been applied across various domains to enhance model development and interpretability, including health data (Pandl et al. 2021), medical imaging (Tang et al. 2021), and the Internet of Things (Shi and Duan 2024). This paper is the first to formally apply data valuation techniques to SIRA.

3 Methods

Let $\mathcal{X}$ be a set of locations where data is collected; $x \in \mathcal{X}$ typically is specified by a longitude and latitude. Let $\mathcal{Y}$ be the set of measurements made over $\mathcal{X}$, here denoting stable isotope ratio values (e.g. $\delta^{13}C$, $\delta^{2}H$, $\delta^{15}N$, $\delta^{18}O$, $\delta^{34}S$), or trace element values (e.g. Si, Cu, S, Ba, Rb). We denote f and g as functions of interest (defined below). For evaluation purposes, we split our data into training and test datasets, where $D = \{z_i\}_{i=1}^{N}$ represents the training dataset with N data points and $z_i = (x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$. Similarly, $T = \{z_i\}_{i=1}^{M}$ denotes the test dataset with M data points. We let $v(h, A, B)$ denote the performance of the model h trained on a dataset A and evaluated on the dataset B, where v would return a numerical value, $v(h, A, B) \in \mathbb{R}$. In cases when the function and the test dataset are known, we drop the dependence in the notation to simply say $v(A)$.

3.1 Forward and Backward Models

Forward Models. For the forward model, the task is to predict the stable isotope values for a given location, i.e., $f : \mathcal{X} \to \mathcal{Y}$. The motivation is to verify whether the characteristics of the location (denoting specified harvest origin) would align with general isotopic values associated with the specified location. Such models can be based on approahes like decision trees, random forests, or XGBoost. Recent works have proposed using Gaussian process regression models with high performance (Truszkowski et al. 2025; Mortier et al. 2024).

Backward Models. In the backward case, we aim to identify the location given measured stable isotope values. We model this relation as $g : \mathcal{Y} \to \mathcal{X}$. The fitted model g would help identify whether the declared harvest location of a species sample aligns with the predicted location from g. Similarly, here, one can use a range of machine learning models to fit this function. For instance, Mortier et al. (2024) reversed a fitted Gaussian process regression model using the Bayes' rule to predict locations from measured isotope ratios.

Atmospheric Variables. In addition, to support either forward or backward models, we often have available a range of atmospheric variables associated with locations. Such variables can be used as either additional inputs to a forward model or auxiliary information in a backward model.

Gaussian Process Regression Models. Gaussian process regression (GPR) models, as explored by (Truszkowski et al. 2025; Mortier et al. 2024), offer a powerful approach for both forward and backward modeling settings. For the forward model $f : \mathcal{X} \to \mathcal{Y}$, GPR can be used to predict isotope ratios $y \in \mathcal{Y}$ at a given location $x \in \mathcal{X}$. We can construct "isoscapes" by fitting independent GP regression models to each feature in $\mathcal{Y}$. Considering our training dataset $D = \{z_i\}_{i=1}^N = \{(x_i, y_i)\}_{i=1}^N$, for each feature $y_j \in \mathcal{Y}$, a GP model is trained to predict isotope values at a new location $x^* \in \mathcal{X}$. The predicted distribution for the feature y_j at location x^* is Gaussian, with a mean and variance given by:

$$\begin{aligned} E[y_j|x^*, \mathbf{X}] =& \mu_j + \mathbf{k}^{(j)T}(\mathbf{K}^{(j)} + \sigma_j^2\mathbf{I})^{-1}(\mathbf{y}_j - \mu_j) \\ V(y_j|x^*, \mathbf{X}) =& k^{(j)}(\mathbf{x}^*, \mathbf{x}^*) + \sigma_j^2 \\ & - \mathbf{k}^{(j)T}(\mathbf{K}^{(j)} + \sigma_j^2\mathbf{I})^{-1}\mathbf{k}^{(j)} \end{aligned}$$

Here, $\mathbf{X} = \{x_1, \ldots, x_N\}$ represents the training locations and $\mathbf{y}_j = [y_{1j}, \ldots, y_{Nj}]^T$ are the values of the j-th feature in the training set. μ_j is the baseline mean for feature y_j, $\mathbf{K}^{(j)}$ is the covariance matrix evaluated at all pairs of training locations, $\mathbf{k}^{(j)}$ is the covariance vector between the test location x^* and the training locations, $k^{(j)}(x^*, x^*)$ is the covariance of x^* with itself, and σ_j^2 is the noise variance for feature y_j. For the backward model $g : \mathcal{Y} \to \mathcal{X}$, we leverage Bayesian inference to reverse the prediction. Given a set of features $y^* \in \mathcal{Y}$ from a location of unknown origin, the posterior probability of its origin being location $x^* \in \mathcal{X}$ is calculated using Bayes' theorem:

$$p(x^*|y^*, D) = \frac{p(y^*|x^*, D)p(x^*)}{\int_{x \in \mathcal{X}} p(y^*|x, D)p(x)dx}$$

The likelihood $p(y^*|x^*, D)$ is derived from the forward GP model, assuming independence of features and using the predicted Gaussian distributions:

$$\begin{aligned} p(y^* \mid x^*, D) = & \prod_{j \in \mathcal{Y}} \frac{1}{\sqrt{2\pi V(y_j \mid x^*, D)}} \\ & \times \exp\left(-\frac{(y_j^* - E[y_j \mid x^*, D])^2}{2V(y_j \mid x^*, D)}\right) \end{aligned}$$

The prior $p(x^*)$ can incorporate prior knowledge about the distribution of tree harvest locations. This Bayesian approach provides a posterior probability map over $\mathcal{X}$, indicating the most likely origin locations for an observation with features y^*. The performance of both forward and backward GPR models can be assessed using the metric $v(h, D, T)$, where h is the GPR model (f or g).

Performance Metrics. The primary metric we will employ is:

$$\text{RMSE} = \sqrt{\int_{\mathbf{x} \in A} \big(d(\mathbf{x}_t, \mathbf{x})\big)^2 \, p(\mathbf{x}|\mathbf{y}^*, D)\, d\mathbf{x}},$$

where $d(\mathbf{x}_t, \mathbf{x})$ is the great circle distance. Comparing RMSE across different GP models helps identify which model minimizes large prediction errors and provides overall reliable estimates.

3.2 Data Valuation

The Shapley value, introduced by (Shapley 1953), offers a principled approach to quantify data value, identifying both highly informative and potentially detrimental data points. The Shapley value ϕ_i for a data point i is computed as the weighted average of its marginal contribution to model performance across all possible subsets of the training data:

$$\phi_i = \sum_{S \subseteq D \setminus \{z_i\}} \frac{|S|!(|D| - |S| - 1)!}{|D|!}[v(S \cup \{z_i\}) - v(S)].$$

Here, D is the full training set, S is a subset excluding i, and $v(S)$ is the model performance (e.g., negative mean absolute error) when trained on subset S. High positive Shapley values indicate highly valuable data points that significantly improve performance, while low or negative values suggest redundancy or detrimental effects, possibly due to outliers, measurement errors, or model misspecification. The Shapley value is not an arbitrary metric; it is uniquely characterized as that satisfying a set of desirable axioms, ensuring fairness and consistency in data valuation:

1. **Efficiency:** The sum of the Shapley values for all data points equals the difference in performance between the model trained on the full dataset and the model trained on an empty dataset: $\sum_{i \in D} \phi_i = v(D) - v(\emptyset)$. This means the total value is fully distributed among the data points.
2. **Symmetry (or Null Player):** If a data point i has zero marginal contribution to every possible subset (i.e., $v(S \cup \{i\}) = v(S)$ for all S), then its Shapley value is zero: $\phi_i = 0$. Useless data points receive zero value.
3. **Linearity:** If the performance metric v is a linear combination of two other performance metrics, $v = a \cdot v_1 + b \cdot v_2$, then the Shapley values for v are the same linear combination of the Shapley values for v_1 and v_2. This ensures consistency across different performance measures.
4. **Dummy:** if two data points i and j always have the same marginal contribution to every subset of D then their shapely value must be equal. $\phi_i = \phi_j$.

These axioms provide a strong theoretical justification for using the Shapley value. Furthermore, the Shapley value can be equivalently expressed as a sum over permutations of the dataset (Shapley 1953):

$$\phi_i = \sum_{\pi \sim \Pi(D)} \left[v(P_i^\pi \cup z_i) - v(P_i^\pi)\right],$$

where $\Pi(D)$ is the set of all permutations of data points in D, π is a permutation sampled uniformly at random from

$\Pi(D)$, and P_i^π is the set of data points preceding instance z_i in permutation π. This permutation-based form is equivalent to the subset definition and underlies the Monte Carlo approximation (Ghorbani and Zou 2019; Kwon and Zou 2021). In practice, we use the permutation form primarily for intuition, while our actual computations employ the subset-sampling TMC-Shapley method.

Truncated Monte Carlo Shapley Value. Because the exact computation of Shapley values is computationally prohibitive, (Ghorbani and Zou 2019) propose their approximation by randomly sampling a limited number of subsets instead of exhaustively considering all possibilities (see Algorithm 1 in the extended version). The key idea is that each random permutation provides an unbiased estimate of the marginal contribution of each data point, convergence achievable in practice in $O(n)$ permutations (typically around $3n$) for an n-point dataset. Generally, Monte Carlo estimators exhibit variance that decreases proportional to $1/\sqrt{m}$ as the number of samples m increases. However, in SIRA settings, each marginal evaluation entails retraining a Gaussian process with $\mathcal{O}(N^3)$ time complexity, creating a pronounced trade-off between computational feasibility and valuation precision. This tension motivates careful calibration of the permutation budget to balance estimator fidelity against real-world processing constraints.

Iterative Data Selection with Shapley Values. To strategically select a subset $D' \subseteq D$ from the original training data D that maximizes model accuracy for both forward and backward prediction tasks, we propose to leverage Shapley values computed once on the full dataset to identify data point importance. We hypothesize that by using these initial global valuations, we can efficiently identify and sequentially remove less valuable samples. Our data selection methodology consists of three distinct steps. Initially, we compute data values for all points in D using the entire dataset. Subsequently, we sort the data points and remove the least valuable one according to these pre-calculated values. Finally, we evaluate the model's performance using the test dataset. This sequential removal process continues as long as model performance improves, relying on the initial single valuation (see Algorithm 2 in the extended version).

Beta Shapley. Building upon the foundational principles of the Shapley value, Beta Shapley offers a flexible generalization for data valuation by recognizing that the standard Shapley value's uniform weighting of a data point's marginal contribution across all subset sizes is not always optimal. In many data valuation tasks, the primary objective is to rank data points rather than precisely distribute the total model performance gain, which makes the strict efficiency axiom less critical. Beta Shapley therefore relaxes this axiom and introduces a weighted-average framework in which weights are governed by a Beta distribution. This enables a more fine-grained valuation by assigning different levels of importance to a data point's marginal contribution depending on the cardinality of the subset it is added to. For instance, by selecting appropriate parameters for the Beta distribution, one can prioritize the contributions made to smaller subsets (see Algorithm 3 in the extended version). We include Beta Shapley primarily to provide conceptual background within the broader family of Shapley-based data valuation methods. Our focus remains on TMC-Shapley, given its superior scalability and lower computational variance in practice.

4 Experiments

We utilized data from two datasets of the genus *Quercus*, collected from various regions worldwide. Two broad datasets were combined in this study with the first dataset comprising tree samples distributed globally ($N = 491$), while the second dataset was focuses specifically on European countries ($N = 287$). Stable isotope ratio measurements were performed following the protocols outlined in (Watkinson et al. 2020; Boner et al. 2007). Our experiments are aimed at answering the below questions:

1. RQ1: What is the role of data Shapley values in the domain of SIRA? (Section 4.1)
2. RQ2: How does model architecture influence data Shapley values and the performance of the proposed data valuation framework? (Section 4.2)
3. RQ3: How does Shapley-based data selection compare against naive or exhaustive baselines? (Section 4.3)
4. RQ4: Can the proposed data valuation framework enhance outcomes in cases involving data imputation for missing or noisy data? (Section 4.4)
5. RQ5: Can data selection methods based on data valuation improve the performance of both directions in SIRA? (Section 4.5)
6. RQ6: Given a specific model and data valuation framework, what level of granularity is optimal for effective data selection? (Section 4.6)
7. RQ7: Do different genera and species within the dataset exhibit varying data Shapley values? Can we identify the most and least important genera or species within the dataset? (Section 4.7)

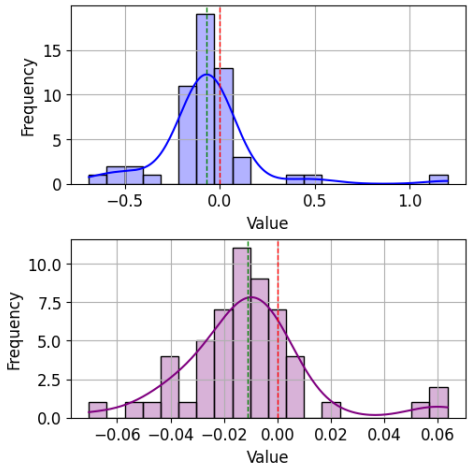
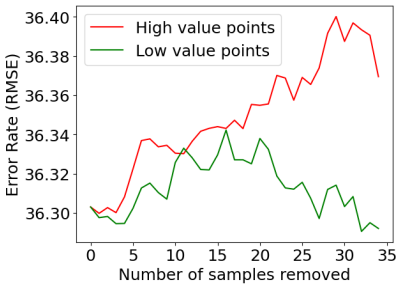
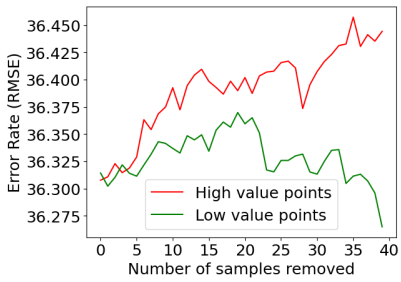


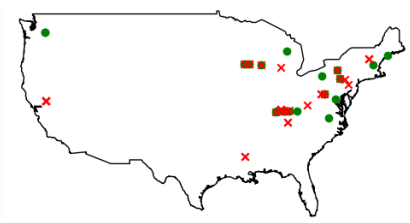
Figure 1: Distribution of data Shapley values, green and red lines represent the mean and the zero, respectively (top: backward model, bottom: forward model); see Section 4.1.



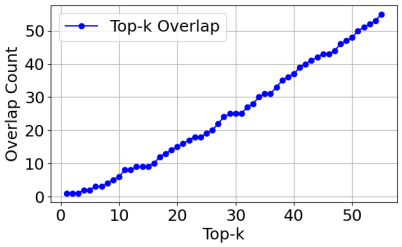
(a) Error plot for GPR



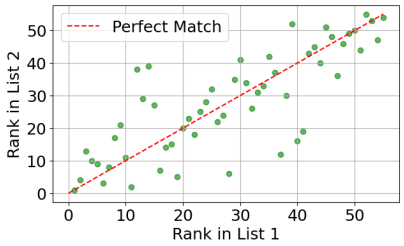
(b) Error plot for RF



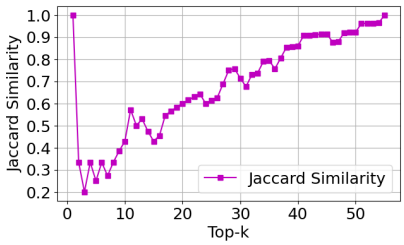
(c) Data selection on a map with Gaussian process regression (perturbed for privacy)



(d) Data Shapley rank overlap with GPR and RF

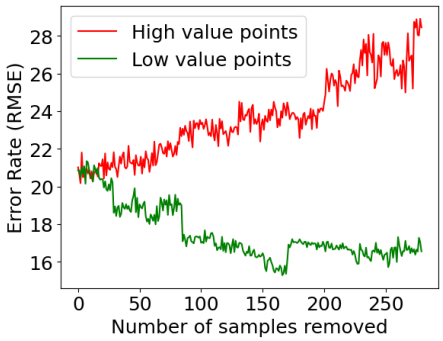


(e) Data Shapley rank scatter-scatter plot with GPR and RF

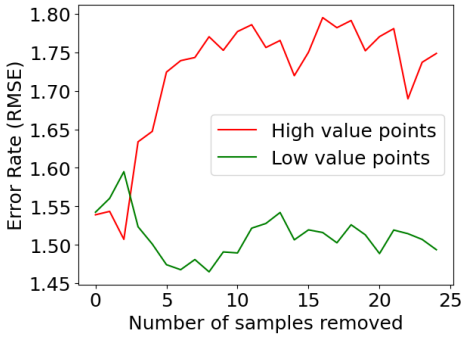


(f) Data Shapley rank Jaccard similarity with GPR and RF

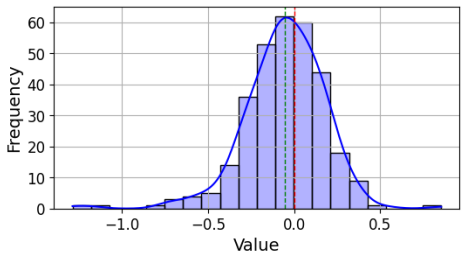
Figure 2: Effect of model architecture on data valuation framework ($N = 87$); see Section 4.2.



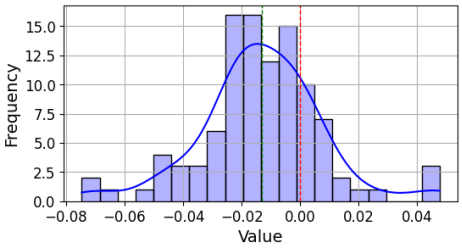
(a) Error plot for median-based imputation



(b) Error plot for listwise deletion



(c) Data Shapley value distribution(median-based imputation)

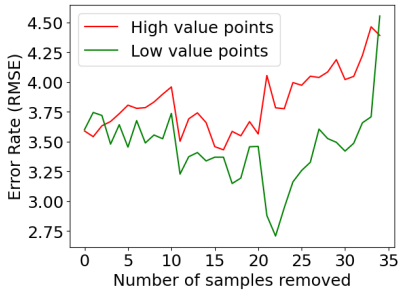


(d) Data Shapley value distribution for listwise deletion

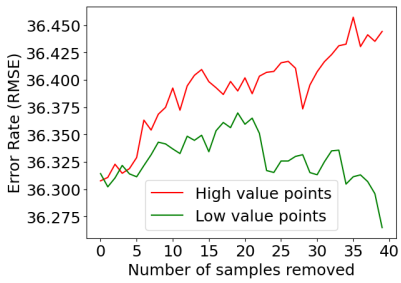
Figure 3: Enhancing missing data strategies through data valuation ($N = 491$); see Section 4.4.

4.1 RQ1: Role of Data Shapley Values in SIRA

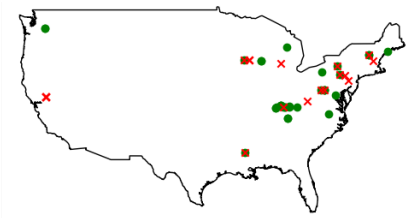
For this experiment, we explore the relevance and application of data Shapley values in the context of SIRA. Fig. 1 presents distribution plots for one subset of the dataset, evaluated in both directions of SIRA analysis, i.e., forward and backward, as described in Section 3. This analysis highlights a significant variation in data Shapley values depending on the dataset and the machine learning model utilized. These



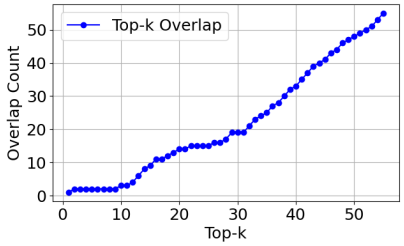
(a) Error plot for backward direction



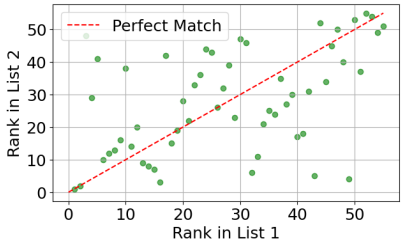
(b) Error plot for random forest on forward direction



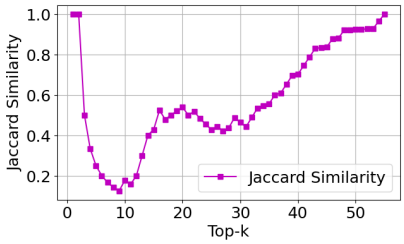
(c) Data selection on USA map with backward direction (perturbed for privacy)



(d) Data Shapley rank overlap for forward vs backward direction

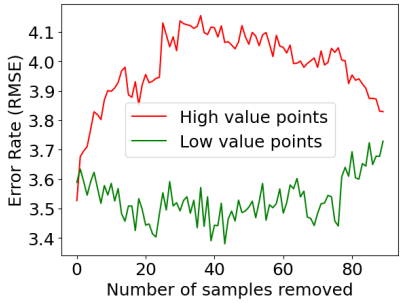


(e) Data Shapley rank scatter-scatter plot for forward vs backward direction

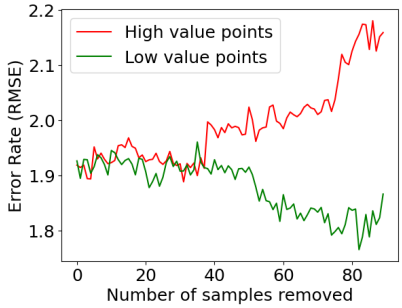


(f) Data Shapley rank Jaccard similarity for forward vs backward direction

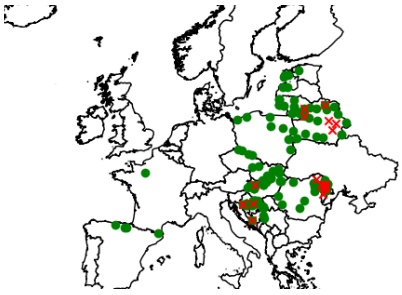
Figure 4: Random Forest-based data valuation on USA only data ($N = 87$); see Section 4.5.



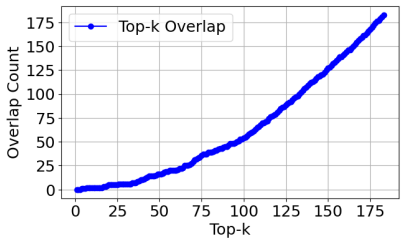
(a) Error plot for backward direction



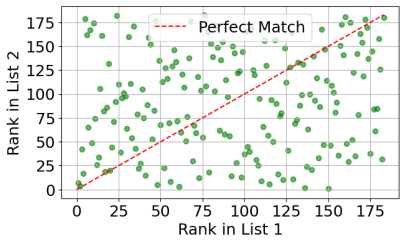
(b) Error plot for random forest on forward direction



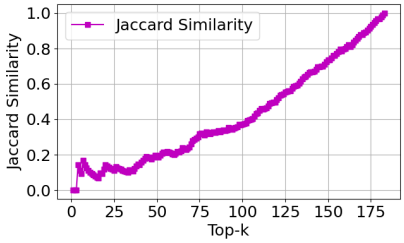
(c) Data selection on the map of Europe with backward direction (perturbed for privacy)



(d) Data Shapley rank overlap for forward vs backward direction



(e) Data Shapley rank scatter-scatter plot for forward vs backward direction



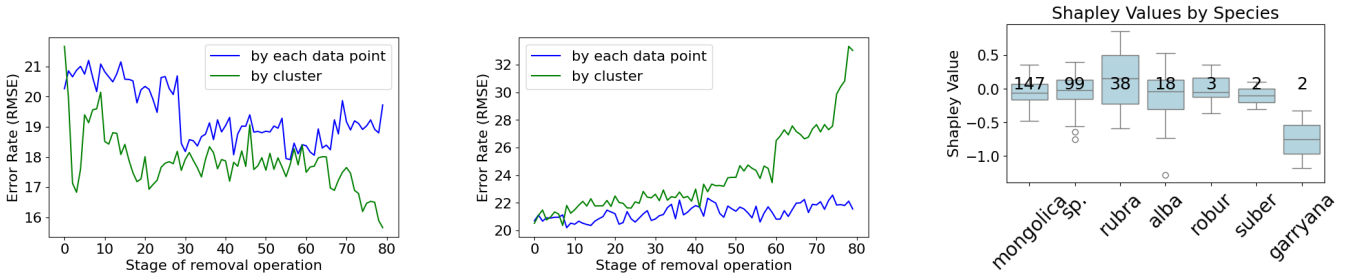
(f) Data Shapley rank Jaccard similarity for forward vs backward direction

Figure 5: Random Forest-based data valuation on Europe only data ($N = 287$); see Section 4.5.

results indicate that subsets with Shapley values exhibiting greater extremities will correspond to larger performance gains following the data selection process.

4.2 RQ2: Influence of Model Architecture

We conducted two types of experiments, removing both high value and low value data points. We present the error plots showing the effect of removing these two types of



(a) Error plot for removing low-value data points

(b) Error plot for removing high-value data points

(c) Data Shapley value distribution per species (Section 4.7)

Figure 6: Optimal Granularity in data selection ($N = 491$); see Section 4.6 and Data Shapley value distribution per species (see Section 4.7).

data points for Gaussian process regression and random forest models. In both architectures, the RMSE increases (indicating performance degradation) when data points with high Shapley values are removed. Conversely, the RMSE decreases (indicating performance improvement) or remains stable when data points with low Shapley values are removed (Figure 2a and Figure 2b). Figure 2c illustrates one such case, where performance improves by removing specific data points (marked as red x on the map). Beyond accuracy, we should point out that random forests, while lacking the theoretically principled uncertainty quantifications of GPR, are much more scalable. We further investigate whether the Shapley values from the two architectures agree with each other. Figures 2(d-f) present three types of rank comparison plots, all of which demonstrate a high degree of agreement between the two models. This result suggests that data value ranks remain consistent across architectures, reinforcing the robustness of our valuation framework even when computational constraints necessitate different modeling choices.

4.3 RQ3: How Does Shapley-Based Data Selection Compare Against Naive or Exhaustive Baselines?

To contextualize the effectiveness of our Shapley-based data valuation framework, we compare it against two natural baselines—random removal of data points and leave-one-out (LOO) removal. Using USA-only data (which already captures the main performance trends observed across our study), Table 1 shows that our Truncated Monte Carlo (TMC) Shapley approach delivers the largest improvement in predictive accuracy, reducing RMSE by 0.8459 compared to the initial model, nearly four times greater than the improvement achieved with random removal and significantly outperforming LOO.

4.4 RQ4: Data Valuation to Support Missing Data Imputation

A critical but often underappreciated issue is that imputation strategies implicitly encode assumptions about isotopic stationarity, e.g., median imputation presumes that the conditional distribution of missing values is homogeneous across all spatial and taxonomic contexts (an assumption that may be violated under latitudinal gradients or localized environmental variation). Similarly, listwise deletion assumes that missingness is independent of the isotopic signal (which is rarely true in practice, as remote sites are both more difficult to sample and may exhibit distinct isotopic patterns). These hidden assumptions complicate the interpretation of downstream model performance and underscore the need for principled data valuation frameworks to mitigate imputation-induced biases. To explore this, we conduct experiments with two common approaches for handling missing data: (i) median imputation, replacing missing values with the dataset median, and (ii) listwise deletion, excluding data points containing missing values from training; in addition, we perform data selection experiments by removing both high-value and low-value data points and examining the resulting performance rankings.

We conducted experiments involving the removal of both high-value and low-value data points. The error plots (Figures 3a and 3b) illustrate the effects of removing these data points under both missing data handling strategies. For both strategies, the RMSE increases (indicating performance degradation) when data points with high Shapley values are removed. Conversely, the RMSE decreases (indicating performance improvement) or remains stable when data points with low Shapley values are removed. We observe significant performance improvements when low-value data points are removed from both strategies, with the improvement being more pronounced for median imputation (26.66%) compared to listwise deletion (5.88%). This result demonstrates the effectiveness of data valuation as a strategy for improving missing data handling. Moreover, the Shapley values for the median imputation strategy exhibit higher magnitudes compared to listwise deletion (Figures 3c and 3d), which aligns with the greater performance improvements observed for median imputation.

4.5 RQ5: Data Selection Methods for Forward and Backward Directions in SIRA

In this experiment, we investigate whether selecting data based on valuation metrics can improve performance in both directions of SIRA. For this analysis, we present the results

Method	Initial RMSE	Best RMSE	Points Removed	Δ RMSE
Random Removal	3.6101	3.4208	2	0.1893
Remove Low-Value with LOO	3.6101	3.2435	24	0.3665
Remove Low-Value with TMC-Shapley (ours)	3.6101	**2.8076**	22	**0.8025**

Table 1: Comparison of data selection strategies. Shapley-based removal consistently outperforms random removal and achieves greater RMSE reduction than LOO, while being more theoretically grounded and computationally efficient. (Section 4.3)

of applying the random forest model in both forward and backward directions using two distinct datasets: USA-only data and Europe-only data.

USA-only data: We present the error plots showing the effect of removing high-value versus low-value data points on the performance of the random forest model. In both directions, the RMSE increases (indicating performance degradation) when data points with high Shapley values, are removed. Similarly, the RMSE either decreases (indicating performance improvement) or remains stable when data points with low Shapley values, are removed (Figures 4a and 4b). Figure 4c illustrates a specific case where performance improves following the removal of certain data points, marked as red x on the map.

We observe performance improvements when low-value data points are removed from the dataset, with the improvement being more pronounced in the backward direction (27.13%) compared to the forward direction (0.14%). This finding demonstrates that performance improvement can indeed vary depending on the direction of SIRA. Moreover, consistent with the observations made in Section 4.4, the Shapley values associated with the backward direction exhibit higher magnitudes compared to the forward direction (Figure 1), which aligns with the greater performance improvements observed for the backward direction.

We further investigate whether the Shapley values from both directions show agreement. To this end, we present three types of rank comparison plots in Figure 4(d-f). All three representations indicate a high degree of agreement between the two directions, suggesting that data value ranks remain relatively consistent across different directions of SIRA. Similar results for the Europe-only dataset are shown in Fig. 5(a-f).

4.6 RQ6: Optimal Granularity in Data Selection

Here, we analyze the appropriate level of granularity for data selection to maximize performance when using a specific model and a data valuation framework. Instead of removing one data point at a time, we consider clusters of fixed distances (in kilometers) and remove all data points within that distance if a data point is selected for removal during the data selection process. We performed this location-based data selection for both high value (Figure 6b) and low value (Figure 6a) data points to understand the impact on model performance. Consistent with the results of previous experiments, we observe performance improvements when low-value data points are removed. Furthermore, the cluster-based removal approach enhances the model's performance more effectively than the one-by-one removal approach.

4.7 RQ7: Species-Specific Data Shapley Values and Their Implications

For this experiment, we explore the variations in data Shapley values across different genera and species, aiming to identify the most and least significant contributors within the dataset. We present a sample distribution plot of data Shapley values for individual species within one dataset (Figure 6c). The plot demonstrates that certain species exhibit significantly higher data Shapley values compared to others, indicating greater contribution to the model's performance.

5 Deployment Details

WFID has operationalized the SIRA approach in this paper to trace the geographic origin of forest products by guiding the collection of high-risk reference samples (Norman 2023), analyzing them in certified laboratories (Fera Science 2025), and deploying trained geographic origin models via the World Forest ID Evaluation Platform (World Forest ID 2024b), which is used by regulators, certification bodies, auditors, and companies to verify sourcing claims across timber and other EU Deforestation Regulation-relevant commodities such as soy, cacao (World Forest ID 2024a), and coffee. In a recent timber market study (Greenfield 2025; World Forest ID 2025), corporate partners used the WFID platform to assess sourcing claims for 59 wood products, focusing on birch due to concerns over sourcing from sanctioned regions, where SIRA and spatial models were used to validate or refute claimed harvest locations. WFID has also supported enforcement efforts, helping identify over 260 tons of allegedly illegal timber (Speed 2024) in Belgium and supporting at least nine additional ongoing investigations.

6 Conclusion and Future Work

Our data valuation methods demonstrate promising results in SIRA analytics, and region-based data selection further enhances performance by removing low-value data clusters. The greatest gains are observed when low-value data points with higher absolute Shapley values are removed, while the largest drops occur when high-value points are excluded, showing that the inferred values are meaningful for product provenance verification. Our analyses further indicate that Shapley-based valuation captures broader spatial and distributional informativeness across regions and species, rather than merely filtering noisy samples, reinforcing its robustness and interpretability for provenance modeling. Future work will generalize these methods across natural resource supply chains, incorporating cross-modal data and exploring greater model scalability and robustness.

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). Ruoxi Jia and the ReDS lab also acknowledge support from the National Science Foundation through grants IIS-2312794, IIS-2313130, and OAC-2239622. 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

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.

Boner, M.; Sommer, T.; Erven, C.; and Förstel, H. 2007. Stable isotopes as a tool to trace back the origin of wood. In *Proceedings of the international workshop "Fingerprinting methods for the identification of timber origins*, 3–5.

Bontempo, L.; Paolini, M.; Franceschi, P.; Ziller, L.; García-González, D. L.; and Camin, F. 2019. Characterisation and attempted differentiation of European and extra-European olive oils using stable isotope ratio analysis. *Food Chemistry*, 276: 782–789.

Camin, F.; Wietzerbin, K.; Cortes, A. B.; Haberhauer, G.; Lees, M.; and Versini, G. 2004. Application of Multielement Stable Isotope Ratio Analysis to the Characterization of French, Italian, and Spanish Cheeses. *Journal of Agricultural and Food Chemistry*, 52(21): 6592–6601.

Covert, I.; Kim, C.; Lee, S.-I.; Zou, J.; and Hashimoto, T. 2024. Stochastic Amortization: A Unified Approach to Accelerate Feature and Data Attribution. *arXiv preprint arXiv:2401.15866*.

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.

Dormontt, E. E.; Boner, M.; Braun, B.; Breulmann, G.; Degen, B.; Espinoza, E.; Gardner, S.; Guillery, P.; Hermanson, J. C.; Koch, G.; et al. 2015. Forensic timber identification: It's time to integrate disciplines to combat illegal logging. *Biological Conservation*, 191: 790–798.

Fera Science. 2025. From complexity to clarity: Why verification is the answer.

Gasson, P. E.; Lancaster, C. A.; Young, R.; Redstone, S.; Miles-Bunch, I. A.; Rees, G.; Guillery, R. P.; Parker-Forney, M.; and Lebow, E. T. 2021. WorldForestID: Addressing the need for standardized wood reference collections to support authentication analysis technologies; a way forward for checking the origin and identity of traded timber. *Plants, People, Planet*, 3(2): 130–141.

Ghorbani, A.; and Zou, J. 2019. Data shapley: Equitable valuation of data for machine learning. In *International conference on machine learning*, 2242–2251. PMLR.

Grant, J.; and Chen, H. K. 2021. Using Wood Forensic Science to Deter Corruption and Illegality in the Timber Trade. https://www.worldwildlife.org/pages/tnrc-topic-brief-using-wood-forensic-science-to-deter-corruption-and-illegality-in-the-timber-trade. Accessed: 2025-02-11.

Greenfield, P. 2025. Sanctioned Russian and Belarusian wood smuggled into UK, study suggests. *The Guardian*.

Jia, R.; Dao, D.; Wang, B.; Hubis, F. A.; Hynes, N.; Gürel, N. M.; Li, B.; Zhang, C.; Song, D.; and Spanos, C. J. 2019. Towards efficient data valuation based on the shapley value. In *The 22nd International Conference on Artificial Intelligence and Statistics*, 1167–1176. PMLR.

Koh, P. W.; and Liang, P. 2017. Understanding black-box predictions via influence functions. In *International conference on machine learning*, 1885–1894. PMLR.

Kwon, Y.; and Zou, J. 2021. Beta shapley: a unified and noise-reduced data valuation framework for machine learning. *arXiv preprint arXiv:2110.14049*.

Liu, H.; Zeng, Y.; Yan, J.; Huang, R.; Zhao, X.; Zheng, X.; Mo, M.; Tan, S.; and Tong, H. 2020. C N H O and mineral element stable isotope ratio analysis for authentication in tea. *Journal of Food Composition and Analysis*, 91: 103513.

May, C. 2017. Transnational Crime and the Developing World.

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.

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.

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

Norman, M. 2023. Tracking Russian birch.

Pandl, K. D.; Feiland, F.; Thiebes, S.; and Sunyaev, A. 2021. Trustworthy machine learning for health care: scalable data valuation with the shapley value. In *Proceedings of the Conference on Health, Inference, and Learning*, 47–57.

Pianezze, S.; Perini, M.; Bontempo, L.; Ziller, L.; and D'Archivio, A. A. 2019. Geographical discrimination of garlic (Allium Sativum L.) based on Stable isotope ratio analysis coupled with statistical methods: The Italian case study. *Food and Chemical Toxicology*, 134: 110862.

Schmitz, N.; Beeckman, H.; Blanc-Jolivet, C.; Boeschoten, L.; Braga, J. W.; Cabezas, J. A.; Chaix, G.; Crameri, S.; Deklerck, V.; Degen, B.; Dormontt, E.; Espinoza, E.; Gasson, P.; Haag, V.; Helmling, S.; Horacek, M.; Koch, G.; Lancaster, C.; Lens, F.; Lowe, A.; Martínez-Jarquín, S.; Nowakowska, J. A.; Olbrich, A.; Paredes-Villanueva, K.; Pastore, T. C.; Ramananantoandro, T.; Razafimahatratra, A. R.; Ravindran, P.; Rees, G.; Soares, L. F.; Tysklind, N.; Vlam, M.; Watkinson, C.; Wheeler, E. A.; Winkler, R.; Widenhoeft, A. C.; Zemke, V. T.; and Zuidema, P. A. 2020. Overview of current

practices in data analysis for wood identification. A guide for the different timber tracking methods.

Shapley, L. S. 1953. A value for n-person games. *Contribution to the Theory of Games*, 2.

Shi, X.; and Duan, H. 2024. Data Valuation and Pricing in Internet of Things: Survey and Vision. In *2024 IEEE International Conference on Smart Internet of Things (SmartIoT)*, 547–554. 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.

Speed, M. 2024. Forest detectives are tackling the illegal wood trade - FT Channels.

Tang, S.; Ghorbani, A.; Yamashita, R.; Rehman, S.; Dunnmon, J. A.; Zou, J.; and Rubin, D. L. 2021. Data valuation for medical imaging using Shapley value and application to a large-scale chest X-ray dataset. *Scientific reports*, 11(1): 8366.

Truszkowski, J.; Maor, R.; Yousuf, R. B.; Biswas, S.; Chater, C.; Gasson, P.; McQueen, S.; Norman, M.; Saunders, J.; Simeone, J.; Ramakrishnan, N.; Antonelli, A.; and Deklerck, V. 2025. A probabilistic approach to estimating timber harvest location. *Ecological Applications*, 35(1): e3077.

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, J. T.; and Jia, R. 2023. A Note on" Towards Efficient Data Valuation Based on the Shapley Value". *arXiv preprint arXiv:2302.11431*.

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.

Wang, T.; and Jia, R. 2022. Data banzhaf: A data valuation framework with maximal robustness to learning stochasticity. *arXiv preprint arXiv:2205.15466*, 19.

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.

World Forest ID. 2024a. Bean to bar: Testing across a cocoa supply chain with ECOCOA. https://worldforestid.org/insights/ecocoa-partnership. Accessed: 2025-07-21.

World Forest ID. 2024b. World Forest ID Evaluation Platforms. https://worldforestid.org/evaluation_platform. Accessed: 2025-07-21.

World Forest ID. 2025. Key findings from timber market study. https://worldforestid.org/insights/key-findings-from-timber-market-study. Accessed: 2025-07-21.

Wu, O.; Zhu, W.; and Li, M. 2024. Is Data Valuation Learnable and Interpretable? *arXiv preprint arXiv:2406.02612*.

Yan, T.; and Procaccia, A. D. 2021. If you like shapley then you'll love the core. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 5751–5759.

Yoon, J.; Arik, S.; and Pfister, T. 2020. Data valuation using reinforcement learning. In *International Conference on Machine Learning*, 10842–10851. PMLR.