LUISS

Degree Program in Data Science and Management

Course of International Operations and Global Supply Chain

Integrating machine learning and compliance requirements of deforestation European directive. The case of global cocoa supply chain.

Prof. Lorenza Morandini

SUPERVISOR

Prof. Alessio Martino

CO-SUPERVISOR

Niloofar Rezaei - 770891

CANDIDATE

Academic Year 2023/2024

Abstract

This thesis explores the application of machine learning techniques for anomaly detection in agricultural supply chains, with a focus on identifying discrepancies in production data and geographical inconsistencies. The research is motivated by the increasing need for sustainable agricultural practices and the enforcement of environmental regulations aimed at preventing deforestation and promoting fair trade practices.

Utilizing a dataset comprising sales and farm size information from Colombian cocoa producers, this study implements machine learning models to detect anomalies in production quantities that are inconsistent with farm capacities and normal production size of each region. Geospatial data analysis is also employed to identify farming activities occurring in non-arable areas or protected regions, which are indicative of potential regulatory non-compliance or environmental harm.

The methodology encompasses data cleaning, integration, and analysis using statistical and machine learning approaches, including clustering algorithms and anomaly detection techniques. The models were trained and validated on historical data, providing a system capable of anomaly detection.

Acknowledgement

I would like to express my deepest gratitude to my supervisor, Professor Lorenza Morandini, for her invaluable guidance, insightful feedback, and unwavering dedication throughout this journey. Her support has been instrumental in shaping this work. I also extend my heartfelt thanks to my co-supervisor, Professor Alessio Martino, for his constructive instructions and valuable contributions, which greatly enriched this research.

A special thanks goes to Alessandro Chelli, Co-Founder and CEO of Trusty, for generously providing the necessary data and offering his expert insights and explanations, which were crucial to this project.

I am profoundly grateful to my boyfriend, Fabrizio Fubelli, for his constant encouragement and support. His thoughtful suggestions and willingness to dedicate time to help me improve my work have been a source of immense strength throughout this process.

Finally, I would like to extend my deepest appreciation to my family, whose unwavering belief in me and their steadfast support, even from afar, have been the foundation that has enabled me to reach this point. Their love and encouragement have truly paved the way for me.

Table of Contents

1. Introduction

In today's globalized economy, the production of commodities such as cocoa, coffee, and oil palm in Africa and South America carries significant economic, social, and environmental implications. These commodities are vital to the economies of several nations, particularly in Africa, where cocoa farming plays a pivotal role in providing livelihoods for millions of smallholder farmers. Africa contributes approximately 77% of the world's cocoa, with West Africa, particularly Ghana and Côte d'Ivoire, being central to this production. The Americas, though smaller in output, account for around 17% of global cocoa production, while Asia and Oceania contribute the remaining 6% (Statista, 2023).

However, the expansion of cocoa cultivation has often come at the cost of rich tropical forests in both Africa and South America, with significant environmental repercussions. This deforestation is linked to biodiversity loss, increased greenhouse gas emissions, and disruptions to local hydrological and soil systems, posing a severe threat to environmental sustainability. In Ghana, for example, the cocoa-chocolate value chain remains central to the economy, yet smallholder farmers, who constitute the backbone of the sector, face numerous challenges, including low incomes and land tenure issues. Despite contributing substantially to national GDP, farmers often live below the poverty line, earning far less than a decent living income (Kwarteng & Emefa, 2023). This economic imbalance perpetuates environmental harm, as farmers are often forced to expand cultivation into forests to increase their yields, leading to further deforestation and land degradation (Kwarteng & Emefa, 2023; Wainaina et al., 2022).

The European Union Deforestation Regulation (EUDR) seeks to address these challenges by mandating sustainable practices within commodity supply chains, including cocoa. By enforcing stricter guidelines for cocoa imports linked to deforestation, the EUDR aims to mitigate the environmental impacts of cocoa production. However, as farmers continue to face economic pressures and limited access to sustainable farming techniques, there is an urgent need for systems that balance environmental conservation with the livelihoods of smallholder farmers (Avadi, 2023).

Smallholder cocoa farmers in Ghana and other key cocoa-producing countries struggle to capture value from the global cocoa-chocolate value chain. Power imbalances and the complex dynamics of the global supply chain make it difficult for these farmers to secure fair prices for their cocoa.

Additionally, factors such as declining soil fertility and shifting rainfall patterns compound these challenges, threatening not only farmers' incomes but also the sustainability of cocoa production itself. Without interventions that provide fair market access and promote sustainable agricultural practices, the livelihoods of these farmers and the ecosystems they depend on remain at risk (Bymolt et al., 2018; Wainaina et al., 2022).
In light of these challenges, it is critical to develop policies and interventions that not only ensure compliance with regulations like the EUDR but also address the underlying economic inequities in the cocoa supply chain. This includes strengthening certification systems and the role of certification bodies in verifying sustainable practices, as well as implementing targeted support for smallholder farmers to enhance their productivity without exacerbating deforestation.

1.1 European Union Deforestation Regulation

The growing concern about the environmental impact of global commodity production has prompted significant legislative measures from entities such as the European Union. The European Union imports millions of tonnes of raw materials annually, a substantial portion of which contributes to deforestation in regions like South America and Africa. The Food and Agriculture Organization (FAO) notes that 90% of global deforestation is linked to agricultural expansion, much of it driven by the supply chains for bulk commodities like soya, cocoa, and palm oil.
The European Union Deforestation Regulation (EUDR) aims to address these challenges by ensuring that the EU's consumption does not promote deforestation or forest degradation. The EUDR sets a cut-off date of December 31, 2024, meaning that commodities produced on lands deforested after this date are not permitted within EU supply chains. This regulation not only applies to areas illegally deforested but also includes those legally cleared in their country of origin after the cut-off date, indicating a stringent approach that surpasses local legal frameworks(European Commission, 2023).

To ensure strict compliance, the EUDR includes substantial penalties for companies that fail to adhere to its mandates. Companies found contributing to deforestation face fines up to 4% of their annual EU turnover, a significant deterrent aimed at ensuring corporate responsibility.

Moreover, non-compliant companies may also face exclusion from public tenders or bans from marketing their products within the EU. These measures are intended to prevent products linked to deforestation from entering the European market, thereby promoting more sustainable production practices globally.

1.2 Key Obligations and Mechanisms of the EUDR

Due Diligence and Transparency: Companies must undertake comprehensive due diligence to verify that their products do not originate from recently deforested or degraded lands. This involves meticulous supply chain mapping to the plot level, with mandatory geolocation data for precise tracing. Businesses are required to maintain records of sourcing locations and production methods for a minimum of five years, enhancing the transparency and accountability of their operations.

Traceability and Risk Assessment: The EUDR mandates traceability of all commodities to ensure they do not contribute to deforestation. This includes keeping accurate records linking products to their origin and producers. Companies must assess and manage the risks of deforestation in their supply chains, implementing measures to reduce these risks to negligible levels.

Legal and Sustainability Compliance: In addition to ensuring no deforestation, companies must demonstrate compliance with the broader legal and sustainability standards of the EUDR. This includes regular environmental impact assessments and supporting farmers to meet these regulations through training and resource provision.

Reporting and Remediation: Businesses are obliged to report any non-compliance discovered within their supply chains to the relevant authorities promptly, ensuring swift corrective measures. This promotes a proactive approach to maintaining deforestation-free supply chains.

Consumer and Authority Information Provision: The EUDR compels businesses to disclose information about their supply chains, including environmental impacts, to both consumers and regulatory authorities. This requirement not only aids consumers in making informed choices but also facilitates regulatory oversight and compliance enforcement.

1.3 Aims and Approach of the Study

With the growing demand for commodities like cocoa, ensuring sustainable supply chains has become increasingly complex. Companies like Trusty, in collaboration with certification bodies, work to verify compliance with environmental regulations such as the European Union Deforestation Regulation (EUDR).

The Trusty platform operates as a blockchain-driven marketplace designed to ensure compliance with new European regulations such as the European Union Deforestation Regulation (EUDR) and the Corporate Sustainability Due Diligence (CSDD). Trusty plays a crucial role in creating a transparent and ethical cocoa supply chain by directly connecting responsible producers with buyers while offering tools for micro-financing and sustainability certification.

Trusty's approach centers on verifying producers' compliance with environmental and social standards through rigorous traceability and certification methods. These efforts help ensure that cocoa products meet EUDR standards, which require geolocalization of cocoa plots, ongoing satellite monitoring to prevent deforestation, and detailed supply chain traceability from farm to market. Trusty supports cocoa producers in maintaining compliance with EUDR by offering data collection tools.

Furthermore, Trusty empowers farmers with access to financing and markets, helping them adapt to sustainability requirements without compromising their livelihoods. Their platform is designed to help smallholder farmers ensure that their products meet the stringent requirements of European markets while fostering a more equitable and sustainable cocoa industry.

Trusty’s integration of blockchain technology guarantees that all data related to the supply chain, including product origin and compliance, remains secure and transparent for buyers, certification bodies, and regulators.

However, the vast amount of data and the large number of farmers make it difficult to manually check every operation. This leads to inefficiencies and the potential for misreporting or oversight.

To address this challenge, a system could help by identifying irregularities in agricultural data, such as discrepancies in reported yields or farming activities in non-agricultural areas. By flagging potential issues, anomaly detection provides targeted insights, helping certification bodies and

companies focus their resources on high-risk areas, making the verification process more efficient and reliable.

Given this backdrop, this thesis aims to develop and implement a framework tailored for detecting anomalies within agricultural data. By focusing on geospatial inconsistencies and production anomalies, the research will contribute to enhancing transparency and compliance in agricultural supply chains, particularly in alignment with the EUDR. This initiative is not only timely but essential, considering the pressing need to address the environmental externalities, including deforestation and its cascading effects on climate and ecosystem. In the following chapters, this thesis will delve into the methodologies employed in crafting machine learning and geospatial models capable of identifying anomalous patterns in agricultural data, thereby supporting the enforcement of the EUDR and contributing to the discourse on sustainable agricultural practices within cocoa-producing regions of Africa and South America.

1.4 Focus Areas of Anomaly Detection

In addressing the challenges of sustainability and regulatory compliance within agricultural supply chains, this thesis identifies and analyzes specific anomalies that compromise environmental standards and economic viability. The focus will be structured into two main categories: **Geospatial Inconsistencies** and **Production Anomalies**.

Production Anomalies

This category focuses on the viability and legality of the reported agricultural output, which is crucial for ensuring that production practices are sustainable and aligned with environmental goals.

1. **Excessive Yield Reporting**: Analyzing cases where the reported production from a given land area exceeds plausible limits based on the country-specific agricultural yield data. Such anomalies can suggest potential inaccuracies or exaggerations in reporting.

Geospatial Inconsistencies

Geospatial data anomalies are critical as they directly impact the integrity of land use and agricultural reporting. The areas of focus include:

1. **Cross-Border Anomalies**: Identification of farming activities reported in coordinates that fall outside the national boundaries, which may indicate errors in data recording or intentional misreporting.

2. **Land Cover Classification**: Detection of agricultural activities reported in locations that are typically non-arable, such as forests, bare lands, water bodies or urban areas. This involves analyzing GPS data to identify discrepancies where farming is claimed but likely infeasible.

3. **Land Change Detection**: Utilizing historical land-use data to determine if deforestation has occurred in areas associated with the farmers in the dataset. This is aligned with the EUDR's requirements to prevent commodity sourcing from recently deforested lands.

1.5 Objectives of Anomaly Detection

The detection of these anomalies aims to:

- **Enhance Supply Chain Integrity**: By ensuring the accuracy and feasibility of the reported data, the research helps in building a more transparent supply chain that stakeholders can trust.
- **Support Regulatory Compliance**: Effective anomaly detection assists in enforcing compliance with environmental regulations such as the EUDR, which is critical for preventing deforestation and promoting sustainable agricultural practices.

By identifying these specific areas of anomalies, this research will contribute significantly to the discourse on sustainable agricultural practices in cocoa-producing regions of Africa and South America. The methodologies to be used for detecting these anomalies will be detailed in the subsequent chapters, providing a robust framework for data analysis and decision-making.

2. Literature Review

2.1 Machine Learning Applications in Supply Chain Management

Machine learning techniques have been increasingly applied to supply chain management to enhance efficiency, predict risks, and detect anomalies across various industries, including agriculture. These techniques can analyze large volumes of data to predict patterns, detect irregularities, and enhance decision-making processes. In agriculture, where sustainability and traceability are key, ML can be applied to monitor crop yields, optimize resource use, and detect potential risks such as environmental impacts or fraud within supply chains.

various ML approaches can be employed in supply chain management, such as predictive analytics, anomaly detection, and optimization models that significantly improve forecasting and planning processes. Incorporating ML enables supply chains to dynamically adapt to changing environmental or economic conditions by accurately predicting future trends and detecting risks early(Tirkolaee & Sadeghi, 2021).

Additionally, by leveraging ML models, agricultural supply chains can dynamically adapt to changes in environmental conditions, such as weather fluctuations or market demands. These models can help mitigate risks by forecasting supply shortages or price volatility, making the entire supply chain more resilient to shocks. In the context of cocoa production, the ability to predict supply disruptions or identify unsustainable practices early can lead to better regulatory compliance and sustainable land use.

Data mining complements machine learning by extracting useful patterns and relationships from large datasets, which is critical for maintaining sustainability across the supply chain. One practical application in this context is the development of pre-warning systems that monitor potential risks related to food safety and sustainability. Rule mining and Internet of Things (IoT) technology can provide real-time tracking of food products, flagging potential risks before they become critical issues. For agricultural products like cocoa, such systems could help monitor various aspects of sustainability, including deforestation risks, water usage, and pesticide levels, ensuring that any deviation from sustainable practices is quickly identified and addressed (Wang & Yue, 2017).

In cocoa supply chains, data mining can be employed to track production

data across multiple regions, analyzing trends such as sudden spikes in yields that may not correspond to the actual land size or capacity. This data-driven approach enables certification bodies and regulators to efficiently identify suspicious patterns in large datasets, providing a targeted method for field inspections and audits.

2.2 Blockchain Technology for Traceability and Compliance

Blockchain technology further enhances the transparency and traceability of agricultural supply chains by providing a secure, immutable ledger of transactions. This technology is especially relevant for commodities like cocoa, where traceability is key to ensuring compliance with environmental standards and regulatory frameworks such as the EUDR. Blockchain records each transaction from farm to consumer, ensuring that every step of the supply chain is verifiable and transparent.

Trusty, illustrates how these technologies are being applied to enhance compliance with the EUDR. By leveraging blockchain, Trusty ensures that every cocoa bean can be traced back to its origin, including details about farming practices, land use, and environmental certifications (Trusty, n.d.).This system allows stakeholders, from regulators to end consumers, to verify that the cocoa they purchase complies with sustainability standards, thus preventing deforestation and land misreporting. Additionally, the integration of blockchain with IoT devices enables real-time monitoring of land use, allowing for immediate detection of activities that may breach regulatory compliance.

2.3 Integration with Geospatial Data and Remote Sensing

In addition to machine learning and blockchain, the use of geospatial data and remote sensing technologies provides another layer of transparency and control in agricultural supply chains. Satellite imagery and geospatial analysis tools can monitor land use changes, detect deforestation in near real-time, and validate that farming practices align with sustainability certifications. By integrating geospatial data with blockchain, supply chains can offer end-to-end visibility, ensuring that commodities like cocoa are produced in compliance with deforestation regulations.

This integration is critical for ensuring that the objectives of the EUDR are met. Platforms like Trusty can combine geospatial data with blockchain to

track deforestation risks, allowing regulators and certification bodies to intervene quickly if anomalies are detected. This technology-driven approach ensures that sustainability in agricultural supply chains is not only a regulatory goal but also a reality, with real-time monitoring and automated compliance checks making it easier to detect and prevent unsustainable practices.

2.4 Anomaly Detection in Supply Chains

The use of machine learning-based anomaly detection systems in collaborative food supply chains has grown significantly in recent years. Study outlines a hybrid anomaly detection framework that combines statistical learning techniques with blockchain technology (Chen et al., 2023). This integration is aimed at improving both the detection of irregularities and enhancing data security across supply chains. The blockchain mechanism ensures that once data is recorded, it is immutable and transparent, reducing the chances of tampering with reported data. Furthermore, the anomaly detection system operates on the principle of identifying data that deviates significantly from expected patterns, whether due to natural fluctuations in supply chain operations or potentially fraudulent activities.

A sophisticated anomaly detection system elaborates on a kernel-based regression model that tracks system performance using sensor data. This model builds on historical observations to predict failures or inefficiencies in real-time operations by identifying abnormal process signals. The framework is designed for continuous monitoring, making it ideal for conditions that evolve dynamically, such as changing agricultural outputs based on weather, soil, or irrigation variations. By creating anomaly bands around the process variables, the system can signal when operations deviate from optimal conditions, thus preventing substantial losses in productivity (An et al., 2011).

Another innovative approach is the pre-warning analysis system, which focuses on early detection and notification of potential risks in the food production supply chain. "As Zhan states, 'the traceability framework monitors various aspects of the production process, including food safety, quality, and environmental impacts' (Ke Zhang et al., 2011). The system uses machine learning to predict potential hazards or deviations." from

standard operations before they occur, offering a safeguard against issues like contamination or unsustainable farming practices. The model analyzes real-time data inputs and historical performance trends to generate alerts, giving supply chain managers ample time to address anomalies.

2.5 Blockchain Integration with Machine Learning

The combination of **machine learning** and **blockchain** technology offers a robust solution for anomaly detection, particularly in ensuring data authenticity and transparency in supply chains. Machine learning models can flag irregular transactions or data points, while blockchain acts as an immutable ledger to prevent tampering. This combination is particularly relevant in agricultural supply chains where ensuring the accuracy and transparency of data related to land use, farming practices, and product quality is critical for regulatory compliance and sustainability efforts. Such systems can also enhance the traceability of products, ensuring that environmental regulations like the EUDR (European Union Deforestation Regulation) are met consistently (Manh et al., 2024)

2.6 Certification's Role in Sustainable Supply Chains

Certification can foster better relationships between stakeholders along the supply chain, including producers, processors, and end consumers. Certification enhances traceability and transparency, thus ensuring that the products marketed to consumers have a verifiable environmental impact. Studies indicate that certification schemes can create market incentives for responsible forest and land use by improving market access and premium pricing for certified products. However, as discussed by Bass et al. (2001), one of the key challenges for certification programs is ensuring equitable access for smallholder farmers, who may find it difficult to meet the financial and administrative burdens of certification (Bass et al., 2001).

2.7 Regulatory Frameworks and Compliance

Regulatory frameworks like the European Union Deforestation Regulation (EUDR) aim to control the importation of commodities linked to deforestation. EUDR requires companies importing products such as cocoa to ensure that their supply chains are free from recent deforestation,

mandating compliance with sustainability requirements. This regulation directly addresses the deforestation problem in cocoa-producing regions by setting cut-off dates for when deforestation must have ceased in areas where commodities are sourced. Companies failing to comply with such frameworks face heavy penalties, as well as potential exclusion from the European market. Such regulations add pressure on certification schemes and companies to verify their compliance with strict sustainability criteria.
Moreover, regulatory frameworks often work hand in hand with certification bodies. For instance, certified farms are better positioned to meet the regulatory requirements of markets like the EU, thus benefitting from smoother trade processes (Bass et al., 2001).

2.8 Challenges in the Cocoa Supply Chain

The cocoa supply chain is fraught with numerous environmental, economic, and socio-political challenges that undermine its sustainability and transparency. While the industry remains a critical source of income for millions of smallholder farmers, primarily in West Africa, Latin America, and Southeast Asia, these regions face persistent obstacles that complicate efforts to enforce responsible agricultural practices and economic equity.
One of the core issues is deforestation, driven by the expansion of cocoa farms into previously forested areas. Deforestation not only leads to the loss of biodiversity but also exacerbates climate change through the release of stored carbon into the atmosphere. Despite the introduction of frameworks like the European Union Deforestation Regulation (EUDR), many regions still struggle to enforce these policies effectively. In particular, smallholders face difficulties in complying with the stringent requirements for sustainability certification, largely due to limited financial and technical resources. Moreover, as noted in the literature, a significant portion of cocoa farms is situated in areas prone to environmental degradation due to weak regulatory oversight (Chunguang et al., 2022).
Another significant challenge revolves around economic instability and income inequality among cocoa farmers. While the global demand for cocoa continues to rise, the economic benefits often do not trickle down to the farmers, who are typically paid low wages for their products. Fluctuating cocoa prices on the world market create further instability, leaving farmers in precarious financial situations. Even when certified as sustainable, the premiums earned from programs such as Fairtrade or

Rainforest Alliance do not always compensate for the cost of compliance, leading to economic stress and the continued reliance on unsustainable practices (Chunguang et al., 2022).

Labor practices are another pressing concern, with widespread reports of child labor and poor working conditions, particularly in West Africa. Regulatory measures have been introduced to combat these abuses, but they are often difficult to enforce in remote farming communities. The reliance on cheap labor exacerbates these issues, and certification schemes, while attempting to address this problem, still have gaps in ensuring full compliance.

From a technological perspective, the lack of traceability and transparency in the cocoa supply chain remains a significant hurdle. Although blockchain and other digital technologies have been proposed as solutions to improve transparency and accountability, their adoption has been slow. Many smallholder farmers lack access to the technology and infrastructure necessary to implement such systems. Consequently, the data provided on the origin of cocoa is often incomplete or unreliable, limiting the ability of stakeholders to verify compliance with environmental and social standards. Anomaly detection, as discussed in the thesis, can play a critical role here, providing companies and certification bodies with insights into irregularities in production data or geographic inconsistencies, offering a potential solution to the traceability problem.

The combination of weak enforcement, economic disparity, labor abuses, and limited technological integration poses significant challenges to achieving a truly sustainable and transparent cocoa supply chain. Certification bodies and regulatory frameworks provide some oversight, but without substantial improvements in governance, technological adoption, and farmer support, these issues are likely to persist.

In conclusion, while regulatory and certification efforts play an important role, there remains an urgent need for innovations in anomaly detection and digital traceability to address the persistent challenges in cocoa production. Certification bodies, together with technological platforms like blockchain, could collaborate more closely to provide better insights into where issues arise, enabling targeted interventions that benefit both environmental sustainability and farmers' livelihoods.

While there has been some research into the applications of machine learning and digital technology in agriculture and supply chain management, particularly with respect to enhancing transparency and

sustainability, there is a noticeable gap when it comes to their specific application in the cocoa supply chain. Studies on supply chain technologies tend to focus on general agricultural systems or large-scale commodities but seldom address the complexities of cocoa production, especially in the context of anomaly detection related to environmental and geographic inconsistencies.

This gap makes this thesis particularly novel. By applying machine learning models such as Isolation Forest for anomaly detection in cocoa production data, and integrating geospatial analysis to track land use and deforestation risks, this work brings a new dimension to supply chain transparency. It introduces methodologies specifically tailored to address the unique challenges of the cocoa industry, such as smallholder farm dynamics, geographic discrepancies, and compliance with regulations like the European Union Deforestation Regulation (EUDR).

Thus, the contribution of this thesis is not only in expanding the body of research on supply chain transparency but also in pioneering a targeted approach that blends machine learning and geospatial analysis to tackle the specific environmental and economic challenges faced by cocoa-producing regions. This creates a foundation for future research that can build upon these methods, making the work both novel and impactful in the broader field of sustainable agriculture.

3.Methodology

3.1 Introduction to Datasets

1.The "Parcels.csv" dataset provided by the Trusty platform encompasses 12,480 records and is an essential resource for analyzing agricultural practices across four diverse countries: Uganda, Togo, Peru, and Ecuador. Each record in the dataset represents an individual agricultural parcel, offering a detailed snapshot of land and farming operations within these geographic and economic contexts.
Geospatial data is provided in the geoJSON format, which includes detailed coordinates and, for some parcels, polygonal boundaries. This data is fundamental for conducting spatial analysis, such as mapping farm locations, analyzing land use patterns, and validating reported information against actual geographical data.

2. The second dataset provided is centered around Colombian farmers and contains detailed information about the farmers themselves. It includes 1,678 records. This dataset is enriched with personal and operational data for each farmer. It includes unique identifiers, which is a personal identification number. Fields like end capture the timestamp of the data entry. with fields such as Map position,' _Map position_latitude', and '_Map position_longitude' and details about the farm size 'areas in cocoa (hectares)'. providing exact coordinates of each farmer's location. This precision is essential for spatial analysis and for correlating agricultural activities with geographical factors.

3. The third dataset in this study focuses on the sales data of Colombian cocoa farmers, comprising a vast array of 89,986 records. This dataset is crucial for evaluating the economic aspects of cocoa production and understanding the market dynamics within Colombia. The dataset provides comprehensive sales data for each transaction, including 'id_association', 'Association Number', which link each sale to specific farmer associations or cooperatives. It encompasses detailed transactional data including the date of sale, quantity of cocoa sold, and financial details like value per kilo and total value of transactions.

3.2 Production anomalies

This study utilizes two latter datasets that provide comprehensive information about Colombian cocoa farmers: 'Farmer_general.csv' and 'Farmer_sales.csv'. Each dataset is pivotal for understanding different aspects of agricultural operations from geographical locations to economic transactions forming the backbone of this anomaly detection.
It is instrumental in correlating financial data with agricultural practices to detect anomalies such as over-reporting of production or sales figures that do not align with the farm size and capacity. Both datasets together offer a holistic view of the farming operations, making them integral to detecting discrepancies that might indicate fraudulent activities or data recording errors.

Libraries Used

Python, being the primary programming language chosen for its robustness and extensive library support, facilitated comprehensive data manipulation and analysis. For data preparation and manipulation, **Pandas** was extensively used, providing powerful data structures and functions for efficient data cleaning, manipulation, and aggregation. **NumPy** supported numerical operations, especially where complex array operations were required.
For the visualization of geospatial data, the **Matplotlib** library alongside **Basemap**, a toolkit extension for Matplotlib, was employed to plot the distribution of farmers and their respective clusters. These visualizations were crucial for providing a geographical context to the anomaly detection results.
The **Scikit-learn** library played a pivotal role, particularly its **Isolation Forest** module, which was used to identify anomalies within the sales and production data. This unsupervised learning algorithm is well-suited for detecting outliers in large datasets, making it an ideal choice for this study. **Scikit-learn** was also used for its implementation of the **K-Means clustering algorithm** and the **silhouette score**, which helped determine the optimal number of clusters for segmenting the farmer data.
To facilitate the clustering and anomaly detection processes, the silhouette scores were calculated to assess the cohesion and separation of clusters, ensuring the optimal grouping of data for subsequent analysis. The

integration of these libraries provided a robust framework for preparing the dataset, executing the clustering, and conducting anomaly detection, with the results being substantiated through comprehensive visual and statistical outputs.
The upcoming sections will delve deeper into the specifics of dataset preprocessing, the application of clustering and anomaly detection techniques, and the fine-tuning of parameters to optimize the models based on the identified patterns and anomalies in the data.

Data Preparation

Preliminary analysis indicated the presence of various data integrity issues common in large-scale agricultural datasets. Several records were missing critical information such as farm coordinates or transaction details, which are essential for any meaningful analysis.
There were instances of misaligned data entries, especially in geographical coordinates and transaction values, likely due to manual data entry errors.
These preliminary findings underscored the need for a robust data cleaning and preprocessing strategy to ensure the accuracy and reliability of the subsequent analyses. All column names were translated from Spanish to English to facilitate easier analysis and understanding.
The first step involved identifying records with missing critical information such as identification numbers, latitude, longitude, and key transaction details. Missing data compromises the ability to perform accurate geospatial and economic analyses.
Rows missing essential geographical information (latitude and longitude) or those lacking crucial sales data (like kilos sold or value per kilo) were removed from the dataset.
For transactions where the total value was not explicitly recorded, it was computed by multiplying kilos sold by value per kilo. This ensured that all records had complete sales data, which was crucial for accurate economic analysis and anomaly detection.
To visually assess and understand the distribution of farmers and identify potential clusters or outliers, the cleaned data set was used to plot the locations of farms across Colombia. This step was essential for recognizing geographical patterns and potential anomalies related to the spatial distribution of cocoa farming operations.

The Basemap toolkit within the Matplotlib library was employed to create geographical scatter plots. This visualization depicted each farmer's location based on their latitude and longitude coordinates. The map provided a visual representation of how farmers are concentrated in Colombia, highlighting areas with dense farming activities and isolated regions which may warrant closer investigation for anomalies or unique agricultural practices.

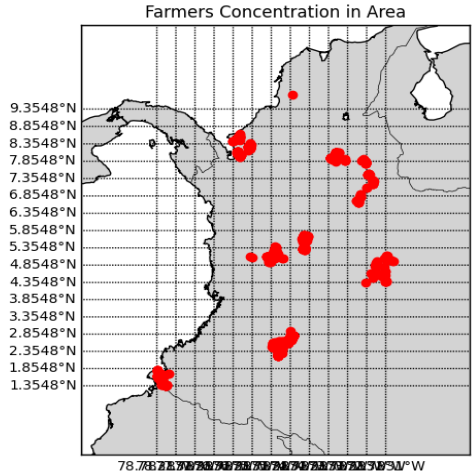

Figure 1: Farmers concentration in area

Clustering was performed to segment the geographical data into manageable groups, facilitating more focused and efficient anomaly detection within localized regions.

Before deciding on the number of clusters, the silhouette score method was applied to determine the optimal clustering arrangement. The silhouette score measures how similar an object is to its own cluster compared to other clusters, providing a clear metric to assess the effectiveness of the clustering. The K-Means algorithm, known for its efficiency in clustering large data sets, was chosen. The algorithm was applied to the latitude and longitude data, and based on the silhouette scores, six clusters were found to be optimal. Each cluster was plotted using a unique color to visually differentiate the groups on the map. This helped in assessing the geographical spread and density of the clusters.

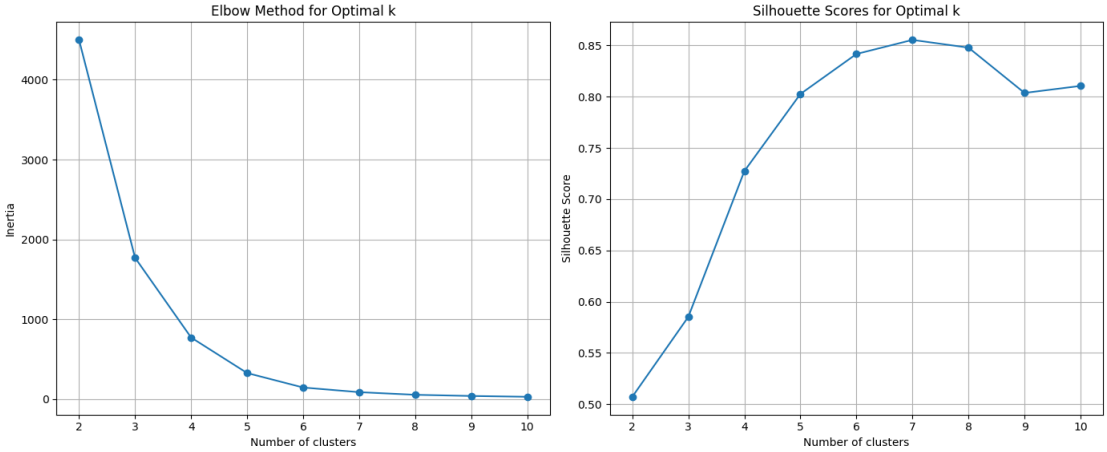

Figure2: Optimal number of clusters

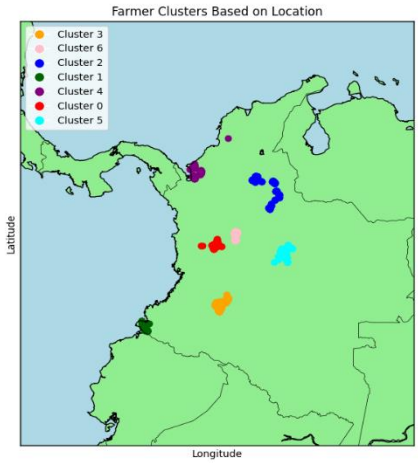

Figure 3: Farmer clusters based on location

3.2.1 Anomaly Detection Approaches

To effectively identify anomalies within the Colombian cocoa farmers' data, two distinct approaches were employed, each utilizing the Isolation Forest algorithm. This algorithm is a type of unsupervised machine learning that is specifically designed to identify anomalies or outliers in data. It works by isolating observations by randomly selecting a feature and then randomly selecting a split value between the maximum and minimum values of the selected feature. This isolation mechanism is particularly effective in datasets with a high dimensionality where anomalies are few

and different. Unlike supervised learning models that require labeled data to learn, the Isolation Forest algorithm operates under unsupervised learning principles. It does not require a training set with labeled outcomes; instead, it identifies anomalies based on the structural properties of the data itself.

In the analysis of sales data from Colombian cocoa farmers, the presence of multiple transaction records for individual farmers presented a unique challenge that necessitated a dual approach to anomaly detection. This dual-method strategy was essential to comprehensively assess both the aggregate behavior of sales over a period and the specific details of each transaction.

Many farmers reported multiple sales transactions. This variability in the number of transactions per farmer allowed for two distinct types of analysis. For farmers with numerous sales records, aggregating these transactions provided a macroscopic view of their total sales activities, which helps in identifying anomalies that could indicate over or under-reporting at a larger scale. Each sales record was also analyzed individually to detect anomalies in specific transactions which might not be apparent when data is aggregated. This approach is particularly important for identifying outliers in transactional data where a single anomalous sale could be obscured by aggregate analysis.

Approach 1: Aggregate Anomaly Detection

For farmers with multiple transactions, the total sales across all transactions were calculated. The Isolation Forest algorithm was then applied to these aggregate figures in relation to the reported hectares of cocoa cultivation to identify outliers. Identifying farmers whose total sales are disproportionately high or low compared to their reported farm sizes and expected production capacities.

Using the Isolation Forest algorithm, anomalies were detected based on the consolidated sales and farm size data. This method helped pinpoint farmers whose total sales metrics significantly deviated from what would be expected given their agricultural capacity.

For this analysis, features such as 'total sales' and 'hectares' were extracted for each farmer. These features were crucial for identifying discrepancies between reported sales and the actual size of the farm. The algorithm was configured with a contamination factor of 0.05, indicating an expectation

that approximately 5% of the data points were outliers. The model was fitted to the data comprising total sales and hectares, allowing it to isolate anomalies based on the statistical properties of the dataset. Farmers flagged by the model as anomalies were marked in the dataset. This marking facilitated further investigation to confirm whether these anomalies were due to data entry errors, misreporting, or possible fraudulent activities.

Approach 2: Individual Transaction Analysis

Each transaction was assessed independently using the Isolation Forest algorithm, considering variables such as kilos sold and value per kilo against the hectares of land farmed. This method helps identify any single transaction that deviates significantly from expected sales patterns based on farm capacity. Evaluate each sales transaction to determine whether the reported sales figures are reasonable based on the hectare of cocoa farmed.

The second approach involved a more granular analysis, where each transaction was evaluated to assess its plausibility based on the corresponding farm size. Transactions for each farmer were individually analyzed. This approach allowed for a detailed assessment of each sale, ensuring that the reported sales volumes were in line with what could realistically be produced based on the farm's size.

Similar to the aggregate analysis, the Isolation Forest was used, focusing on individual transactions' 'total value' and 'hectares'. Each transaction was assessed, with the model identifying those that were statistically unlikely, given the farm size and typical production yields.

3.3 Geospatial Analysis

The initial step in analyzing geospatial anomalies, involved consolidating and cleaning the data collected from various countries. The geographical data extraction targeted retrieving precise locations from the geoJSON structure. For parcels represented by polygons, centroids were calculated to establish a singular, representative coordinate. Direct coordinates were utilized for parcels explicitly defined by points.

Simultaneously, additional relevant attributes such as the parcel's country,

the associated person's ID, and coordination of each farm were extracted. Once coordinates and necessary attributes were extracted, they were integrated back into a main CSV file, aligning all relevant data into a structured format. This file was specifically structured to include essential identifiers, geographical coordinates and country information key elements for subsequent analytical phases.

3.3.1 Cross-Border Anomalies

This section of the methodology focuses on identifying agricultural activities reported in coordinates that fall outside the designated national boundaries. To achieve this, highly accurate geographic boundary data was sourced from Natural Earth, which provides vector data at a 1:10 million scale. This data includes detailed country boundaries necessary for precise geospatial analysis. This dataset is renowned for its accuracy and is widely used in geographic analyses that require precise country boundary delineations. This data has 100% of accuracy as claimed by the data providers.

Libraries Used

GeoPandas library extends the functionalities of pandas to allow spatial operations on geometric types. GeoPandas was crucial for operations such as reading, manipulating, and analyzing geospatial data, which are foundational in processing the shapefiles of country boundaries.
Pandas Used for its robust data structures and tools for data manipulation and analysis. It was especially useful in handling tabular data, merging datasets, and transforming coordinate data into structured formats that could be easily analyzed.
Shapely library was employed for the manipulation and analysis of planar geometric objects. It facilitated the conversion of latitude and longitude coordinates into point objects, which were then used to determine spatial relationships, such as containment within country polygons.

Anomaly Detection Setup

A GeoDataFrame was set up to manage the country boundaries data efficiently. This framework was implemented to load the country

boundaries only once throughout the session, which enhances the performance by avoiding repetitive loading of the same dataset. This method is especially effective in handling large datasets where operations need to be optimized for speed and memory usage.

For the analysis, a reverse geocoding technique was applied. This technique involves converting geographic coordinates (latitude and longitude) into a two-letter ISO country code. This process checks whether the provided geographic coordinates fall within the recognized boundaries of a country according to the shapefile data. This step determines whether the location data accurately corresponds to the reported country, which identifies any discrepancies that might indicate data recording errors or possible intentional misreporting.

Using the reverse geocoding functionality, the study analyzed a dataset containing latitude and longitude of farms from five different countries. Each coordinate was checked against the Natural Earth country boundaries to determine if the location fell outside the reported country's boundaries.

The analysis revealed approximately 59 anomalies, primarily occurring in Uganda and Ecuador. These anomalies are likely due to errors in data recording or intentional misreporting and were flagged for further investigation.

Observations from the Distribution Maps

A small number of anomalies were observed along the borders, suggesting a potential misalignment in the geospatial data entry or a misinterpretation of border definitions. Some of these coordinates lie close enough to the border that they might be explained by the natural imprecision inherent in GPS technology or discrepancies in the mapped boundary data. With coordinates spread significantly inside Uganda, there's an indication of either widespread misreporting or a systematic issue with how coordinates are logged in this region. This could have broader implications for any agricultural policies or economic decisions based on this geospatial data.

Figure

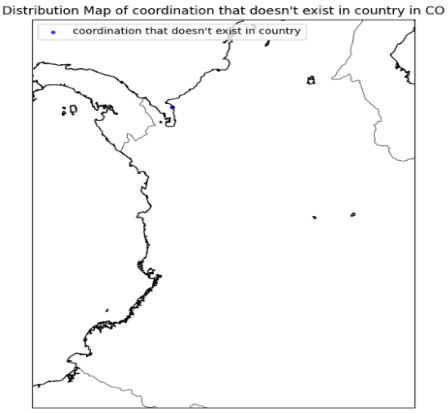

Figure 4: Cross-Border Anomalies in Colombia

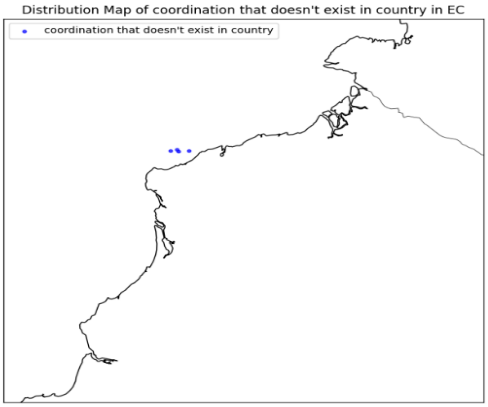

Figure 5: Cross-Border Anomalies in Ecuador

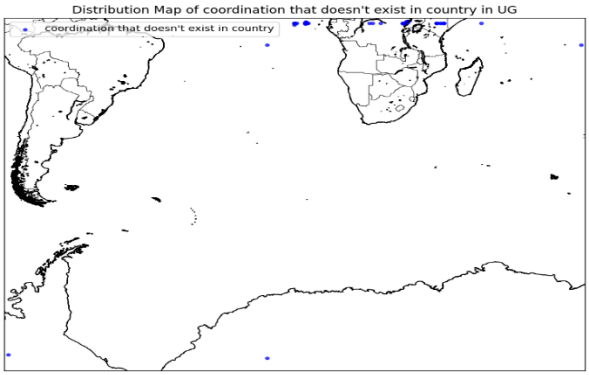

Figure 6: Cross-Border Anomalies in Uganda

3.3.2 Land Cover Classification

For this part, I established specific criteria based on the types of land cover misclassifications that would indicate anomalies. The criteria included checking if reported farm coordinates fell into non-agricultural land cover classes such as urban areas, water bodies, forests, bare areas, or snow-covered regions.
For each farm location, the surrounding land cover type was verified against the expected agricultural designation. Locations categorized under inappropriate land cover types, according to these criteria, were flagged as anomalies. This method allowed for identification of geographical areas where the land use did not align with the reported agricultural activities.

Data Source and Acquisition

1. Copernicus

For the land cover classification component of the anomaly detection in agricultural data, high-quality satellite imagery is essential. To this end, datasets were sourced from the Copernicus program, specifically the land cover data derived from the Sentinel-3 satellite observations for the years 2020, 2021, and 2022(Lamarche & Defourny, 2024). Here are the specifics of each dataset:

2020 Land Cover Data: Approximate accuracy of 70%. This dataset serves as the baseline for observing changes and establishing a normative pattern of land cover that will assist in identifying deviations in subsequent years.
2021 Land Cover Data: Slightly higher accuracy of 70.48%. The slight increase in accuracy helps in refining the analysis and providing more confidence in the anomaly detection process for this year.
2022 Land Cover Data: Accuracy maintained at 70.30%. This consistency in dataset accuracy over the years allows for a reliable comparative analysis across the timeframe.
The datasets are available for public access through the Copernicus Climate Data Store (CDS), providing a resource for researchers and analysts to download and use the data for various environmental monitoring purposes.

The Sentinel-3 series data, categorizes land cover using a hierarchy based on the Land Cover Classification System (LCCS) developed by the UN Food and Agriculture Organization (FAO). This system enables a detailed classification of land cover types that are essential for environmental monitoring and management.

The Copernicus Land Cover (LC) data is a key resource for environmental and land monitoring applications, crafted using comprehensive methodologies to ensure high levels of accuracy and reliability. Here's an overview of the methodologies and accuracy measures employed in the production of the Copernicus LC maps.

Copernicus utilizes satellite imagery primarily from the Sentinel series of satellites, which offer wide coverage and frequent revisits. The data includes both optical and radar images, providing a diverse range of information for land cover classification. The satellite data undergoes several preprocessing steps to correct for atmospheric, angular, and other sensor-specific distortions. This step is crucial to ensure that the data accurately represents the Earth's surface conditions. The program uses advanced machine learning techniques, including deep learning and ensemble classifiers, to categorize the land surface into different land cover types. These classifiers are trained with a large and diverse set of ground-truth data collected from various sources.

Each LC map is subjected to a systematic quality control process. This involves checking the classification results on a regular grid to identify any potential errors such as misclassifications or boundary issues. Detected errors are then corrected in a post-classification refinement step.

Depending on the heterogeneity of the landscape, a variable-sized grid is used to manually inspect and validate the land cover data. This manual validation is crucial for ensuring the accuracy of land cover classifications in diverse environments.

A brief overview of the land cover categories included in the Copernicus data:

1. **Cropland**: Includes both rainfed and irrigated areas. Specific subclasses distinguish between herbaceous cover and tree or shrub cover.
2. **Mosaic Land**: Combines cropland with natural vegetation, either with cropland or natural vegetation being more dominant.

3. **Forest Areas**: Differentiated by leaf type (broadleaved or needleleaved) and the deciduous or evergreen nature of the trees. It also includes mixed leaf type forests.
4. **Shrubland and Grassland**: Includes pure shrubland areas, grasslands, and mosaics of trees, shrubs, and herbaceous cover.
5. **Wetlands**: Classified by tree cover in flooded regions, both in fresh or saline water.
6. **Urban Areas**: Distinguished from natural landscapes, indicating regions of human habitation and infrastructure.
7. **Bare Areas**: Includes areas with little to no vegetation cover, such as deserts and rocky areas.
8. **Water Bodies and Snow**: Covers areas of water, including seas, lakes, and rivers, as well as regions permanently covered by snow and ice.

Libraries used

Xarray is employed to manage the multi-dimensional arrays of land cover data derived from satellite images. It handles the slicing of data by geographical coordinates and supports operations across different dimensions of the data.
Shapely is crucial for creating and manipulating geometric shapes like points, lines, and polygons. It is used to convert latitude and longitude into point objects and perform spatial operations such as calculating distances and checking if a point lies within a polygon.
Spatial Indexing (Part of GeoPandas) is implemented through GeoPandas' sindex property on GeoDataFrames. It is used to quickly locate the nearest land cover classification point to any given agricultural coordinate.

Methodological Framework

The global dataset was quite large, encompassing land cover information for the entire planet. To make the analysis manageable and relevant, I extracted data for specific countries using bounding boxes.
Bounding boxes define the geographical limits (latitude and longitude) of each country. These are used to slice the global dataset to obtain a subset of data that only includes the area within a country's boundaries.
Once the country-specific dataset was sliced from the global data, it was

converted into a GeoDataFrame a data structure optimized for spatial data. A spatial index is created on this GeoDataFrame using GeoPandas, which significantly speeds up spatial queries such as finding the nearest point.
The method involves identifying the nearest land cover data point for each reported agricultural coordinate, rather than using the exact point. This approach is chosen due to the resolution of the satellite data.
Copernicus land cover data is typically provided at a 300-meter resolution. This means each data point represents the predominant land cover classification within a 300-meter square area. As such, exact match queries (looking for the exact latitude and longitude in the dataset) would often fail because the coordinates reported by farmers are unlikely to match the discrete grid points exactly.
Instead, a nearest neighbor search is conducted. This search finds the closest data point within the dataset to the reported coordinate, ensuring that the analysis accurately reflects the land cover type that most closely corresponds to the reported location.

2. ESA World Cover

The other data source that was acquired for the land cover classification part was ESA World Cover data. I did this to ensure higher precision and reliability on the results of the classification.
The ESA WorldCover project offers high-resolution global land cover maps, utilizing the advanced observation capabilities of the Sentinel satellites operated by the European Space Agency (ESA). These datasets are particularly beneficial for studies that require precise and up-to-date information on land use and land cover across different environmental contexts, including agricultural monitoring and environmental impact assessments.
The ESA WorldCover datasets provide land cover maps at a 10-meter resolution. This high level of detail is crucial for accurate land cover classification, especially in diverse and fragmented landscapes where changes in land use might occur in small patches that are not detectable at coarser resolutions.
The ESA WorldCover datasets undergo extensive validation procedures to ensure accuracy. These include comparisons with ground-truth data and independent validation by third-party organizations, ensuring that the

classifications are reliable and can be confidently used.
The ESA WorldCover project employs advanced methodologies to generate land cover maps with significant accuracy. These methods leverage state-of-the-art satellite imaging technology and sophisticated data processing techniques to ensure that the land cover classifications meet the needs of various applications, from climate modeling to biodiversity conservation.
The dataset primarily utilizes multispectral imagery from the Sentinel-2 satellites. This data is characterized by high spatial resolution and multiple spectral bands that are ideal for distinguishing different types of land cover. The raw satellite data undergoes rigorous preprocessing, which includes atmospheric correction to remove the effects of aerosols and other atmospheric conditions. This step ensures that the spectral signatures are as accurate as possible, reflecting the true colors and characteristics of the land surface.
ESA WorldCover uses machine learning algorithms, particularly random forests and support vector machines, to classify the preprocessed imagery into various land cover types. These algorithms are trained on a globally distributed set of training data, which includes ground-truth observations and other satellite data. After classification, the datasets are validated using independent validation datasets not used during the training of the classification models. This process involves statistical methods such as the confusion matrix, which provides detailed insights into the accuracy of the classification.

2020 Dataset: The overall accuracy reported for the 2020 version of the ESA WorldCover dataset was approximately 74.4%. This reflects a high level of reliability but also highlights the challenges in certain complex landscapes or land cover types (Van De Kerchove, 2020).
2021 Dataset: The subsequent year saw a slight improvement in accuracy, reaching approximately 76.7%. This enhancement can be attributed to refinements in the classification algorithms and better training data (Van De Kerchove, 2022).

A brief overview of the land cover categories included in the Copernicus data:

1. **Tree cover**: Areas predominantly covered by trees, often used to describe regions where tree canopy covers a significant portion of the ground.
2. **Shrubland**: Regions covered by woody vegetation shorter than trees, often found in semi-arid areas and used for grazing.
3. **Grassland**: Areas where grasses predominate, typically used for pasture and haylands; these regions may also include herbaceous types of vegetation.
4. **Cropland**: Land primarily used for the cultivation of crops. This includes areas under annual and perennial crops and orchards.
5. **Built-up**: Areas substantially covered by buildings and other man-made structures. This includes cities, towns, villages, and transportation infrastructure.
6. **Bare / sparse vegetation**: Lands with limited or no vegetation, including desert areas, rock, and sand surfaces, and areas of extensive urbanization where little to no vegetation is present.
7. **Snow and Ice**: Regions permanently covered by snow or ice, not subject to significant seasonal variation.
8. **Permanent water bodies**: Bodies of water that do not significantly fluctuate in volume throughout the year, such as lakes, reservoirs, and large rivers.
9. **Herbaceous wetland**: Areas where the soil is saturated with moisture either permanently or seasonally, covered predominantly by herbaceous vegetation.
10. **Mangroves**: Coastal wetlands found in tropical and subtropical regions, characterized by salt-tolerant trees and other plant species adapted to life in saline water conditions.
11. **Moss and lichen**: Areas primarily covered by mosses and lichens, often found in arctic or mountainous environments where conditions do not favor the growth of higher plants.

Libraries Used

Rasterio is a library that allows for the reading and writing of geospatial raster data. Rasterio is employed to open and process raster images from the ESA dataset, specifically for accessing specific land cover class information encoded within GeoTIFF files that represent different regions and years.

Math Library which Provides access to mathematical functions defined by the C standard. In this analysis, it is used primarily for rounding operations, ensuring that coordinates are processed correctly according to the raster grid specifications.

Methodological Framework

The methodology consists of extracting and analyzing land cover classifications for given geographic coordinates, focusing on validating the land type and detecting any potential anomalies.
Coordinates were first rounded and formatted to match the naming conventions of the stored raster files (GeoTIFFs). This ensures the correct raster file corresponding to the geographic location is accessed. Using Rasterio, the GeoTIFF files are opened, and a window of data around the specified coordinates is extracted. This windowed approach allows for analyzing the immediate area around the given point, enhancing the accuracy of the land cover classification check.
The extracted raster data, representing land cover classes in a matrix format, is processed using NumPy to count occurrences of each land cover class within the window. This step is crucial for determining the predominant land type at the location.
Each land cover class is identified by a unique code, which is mapped to a descriptive label using a predefined dictionary (lccs_class_labels). This mapping facilitates understanding and reporting of the land cover types.
Spatial queries are performed to ensure that the data point lies within the correct geographic boundaries as per the ESA dataset. This involves checking if the point falls within the bounds of the raster data and if necessary, adjusting the query to fit the data's spatial resolution.

3.3.3 Selective validation

When evaluating the accuracy and reliability of the land cover classifications obtained from the Copernicus satellite and ESA World Cover data, a practical approach was implemented by manually verifying specific cases through an additional, widely trusted source Google Maps.
To optimize the verification process, coordinates were selectively chosen based on specific criteria such as anomaly flags or areas of particular interest, aiming to scrutinize the most impactful or suspicious data points.

This targeted approach helps in efficiently using resources while enhancing the validity of the data assessment.
Despite these efforts, it is important to acknowledge that only a subset of the data points was verified, leaving the possibility of undetected errors or anomalies within the unverified segments of the dataset. The vast size of the dataset and the labor-intensive nature of manual verification impose practical limits on the scope of validation efforts. To mitigate this limitation and enhance data integrity, it is recommended that regular ground checks be conducted by certification bodies. These checks should be systematic and periodic, focusing on randomly selected coordinates or areas previously identified as problematic, to ensure ongoing accuracy and reliability of the land cover data.

3.3.4 Comparison between Copernicus Data and ESA World Cover

While both ESA WorldCover and Copernicus satellite programs are European initiatives that provide environmental data, there are notable differences in their focus and applications:
Resolution and Detail: While Copernicus also offers high-resolution data, ESA WorldCover's specific focus on 10-meter resolution for land cover provides finer details that are crucial for certain types of environmental and agricultural analysis.

Frequency of Updates: ESA WorldCover's annual updates offer more frequent data refreshes compared to some Copernicus products, which may update less regularly depending on the specific service and dataset.

Methodological Differences: ESA WorldCover uses a unique set of algorithms and validation techniques tailored specifically for land cover classification. These methodologies differ from those used in Copernicus datasets, which can cover a broader range of environmental monitoring applications beyond just land cover.

By combining Copernicus to ESA data, this research benefits from the enhanced resolution and updated methodologies, providing a more detailed and current snapshot of land cover dynamics. This transition is critical in identifying subtle yet significant changes that might be overlooked by coarser, less frequently updated datasets.

3.3.5 Geospatial Anomaly Detection Setup

To ensure the accuracy and reliability of land cover classifications in agricultural datasets, I integrated and compared data from both the Copernicus and ESA datasets. By cross-referencing these sources, the research aims to identify discrepancies that may signal anomalies such as misreported or misclassified land use.

The datasets from Copernicus and ESA for the years (2022 from Copernicus) 2021 and 2020 were meticulously prepared for analysis. This involved mapping the raw classification data to a unified set of categories to facilitate comparison. The classifications include various land types such as 'Agriculture', 'Forest', 'Settlement', and 'Water'. Each land cover classification was translated into these broader categories to standardize the data across different sources.

The anomaly detection was structured around three primary criteria:

- **Cross-Dataset Discrepancies:** Any significant variation in land cover classifications for the same geographic coordinates across the Copernicus and ESA datasets was flagged. These discrepancies may indicate potential errors or falsifications in the reported data.
- **Non-harvestable Land Cover:** Areas classified as non-agricultural such as urban regions, water bodies, and barren lands were flagged as 'non-harvestable'. This category highlights regions wrongly reported as agricultural, pointing towards possible misreporting.
- **Cross-Border Anomalies:** Coordinates that do not match their reported national boundaries were identified and flagged, suggesting inaccuracies in geolocation data.

3.3.6 Implementation of Detection Algorithms

The detection algorithms were implemented using Python, with libraries such as Pandas for data manipulation and Rasterio for handling geographical data. Specific functions were developed to map land cover classifications and flag anomalies. Each land cover category from the datasets was mapped to predefined broad categories then custom functions were written to systematically check each row in the dataset against the anomaly detection criteria. Rows meeting any criteria were marked as potential anomalies.

To visually analyze the distribution of anomalies, geographical maps were

generated using Matplotlib and Basemap. These maps provide a visual representation of where anomalies are concentrated, particularly highlighting:

Anomaly Distribution by Country: Maps illustrating the spatial distribution of anomalies across different countries, aiding in the identification of regions with frequent discrepancies.

Comparison of Anomalies Across Classifications: Separate maps for each type of anomaly (cross-dataset discrepancies, non-harvestable land cover, and geographical inconsistencies) to visually assess the patterns and extent of the issues detected.

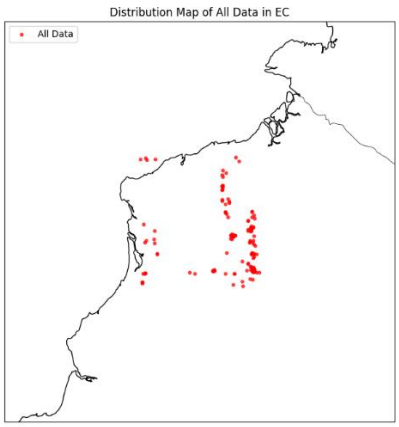

Figure 7: Gesopatial Anomalies in Ecuador

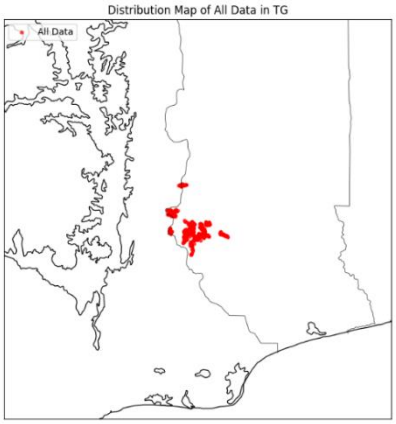

Figure 8: Gesopatial Anomalies in Togo

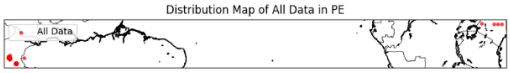

Figure 9: Gesopatial Anomalies in Peru

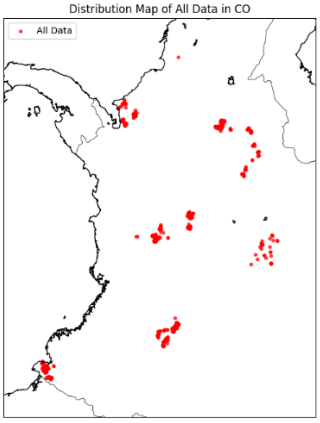

Figure 10: Gesopatial Anomalies in Colombia

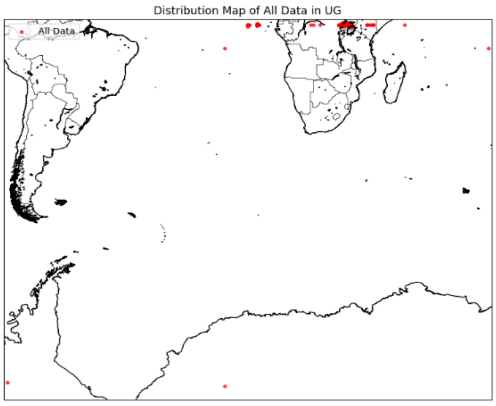

Figure 11: Gesopatial Anomalies in Uganda

3.3.7 Land Cover Changes

This segment of the research focuses on the detection of land cover changes over time at specific geographic locations, utilizing a robust analysis of sequential land cover classifications. This method identifies shifts in land use by comparing the land cover data from one year to subsequent years, providing insights into agricultural dynamics and potential misreporting or misclassification that leads to deforestation.
The analysis of land cover changes relies on a systematic approach to processing spatial and temporal data derived from NetCDF files spanning the years 2018 to 2022. This section delineates the methodology employed to handle the data, especially addressing the challenges posed by the temporal mismatch between the land cover data availability and the extended timeline of farmer data up to 2024.

Data Sources and Preparation

The data for this analysis was sourced from comprehensive land cover datasets, specifically formatted NetCDF files that contain yearly snapshots of land classifications across multiple geographic regions. Initial data from various sources, was cleaned and standardized. Columns were renamed for consistency, and essential attributes such as latitude, longitude, and area were extracted. Coordinates were derived from the GeoJSON fields, ensuring that each land parcel's centroid was accurately calculated for

point data or appropriately estimated for polygons.
The primary function, is designed to extract land cover data corresponding to specific geographic coordinates (latitude and longitude) of each farmer's parcel. This extraction is constrained by the temporal coverage of the available NetCDF files, which document land cover from 2018 to 2022. For each farmer's data entry, the function identifies the closest available land cover data point, typically defaulting to the latest available year, 2022, when encountering data entries for 2023 or 2024.
Given that the land cover datasets extend only up to 2022, any farmer data from 2023 or 2024 inherently lacks corresponding land cover information. The script manages this by applying the land cover data from the latest available year (2022) to these future entries then utilizing the temporal marker from the last available dataset to annotate these data points, thereby indicating that these are extrapolations rather than observations from those specific years.

Land Cover Categorization

A custom function (categorize_land_cover) was employed to map raw classification codes to recognizable land cover types such as Agriculture, Forest and Settlement. This categorization aids in the comparative analysis across different datasets and over multiple years.
The script incorporates a robust mechanism for detecting changes in land cover over the available time span through the following steps:

1. **Extraction per Time Point:** For each time point available in the NetCDF files, the script fetches land cover data for the relevant grid cells that align with the farmer's parcel coordinates.
2. **Sequential Comparison:** The check_land_cover_changes function orchestrates a year-by-year comparison of land cover classifications. It sorts the data chronologically for each unique parcel defined by farmer ID and coordinates and checks for any alterations in the land cover type from one year to the next.
3. **Change Flagging:** Any transition in land cover type, such as a shift from 'forest' to 'agriculture', is flagged as a significant change. This flagging is critical for identifying substantial modifications in land use which might impact agricultural practices and environmental assessments.

Change Detection Process

The core of this analysis is the 'process_farmers_land_cover' function, which:

Intersects Farmer Data with Grid Cells: Each farmer's land parcel is aligned with the corresponding grid cell within the NetCDF dataset. This step considers the area of the parcel in relation to the standard grid size (9 hectares per grid cell) to ensure comprehensive coverage.
Sequential Comparison: The land cover classification for each grid cell is tracked over consecutive years. Any shift in classification from one year to the next is flagged as a change.
Spatial and Temporal Granularity: The process accounts for spatial granularity by examining multiple grid cells covered by larger parcels and temporal granularity by analyzing changes over each available year.
Data Recording: All detected changes are recorded, noting the previous and subsequent land cover classifications.
Using the Basemap toolkit, geographical maps are generated to visually represent the location and extent of land cover changes. These maps display the 'before' and 'after' states of land changes, providing a clear visual representation of transitions over time.
For illustrative purposes, consider a farmer's parcel that is geographically constant over five years with the following land cover transitions:

- **2018:** Forest
- **2019:** Forest
- **2020:** Agriculture
- **2021:** Agriculture
- **2022:** Agriculture

The methodology will identify a significant change in land cover between 2019 and 2020, marking the transition from forest to agriculture. This change, once detected, is documented and analyzed for its implications on land use and agricultural sustainability.
The methodology will identify a significant change in land cover between 2019 and 2020, marking the transition from forest to agriculture. This change, once detected, is documented and analyzed for its implications on land use and agricultural sustainability.

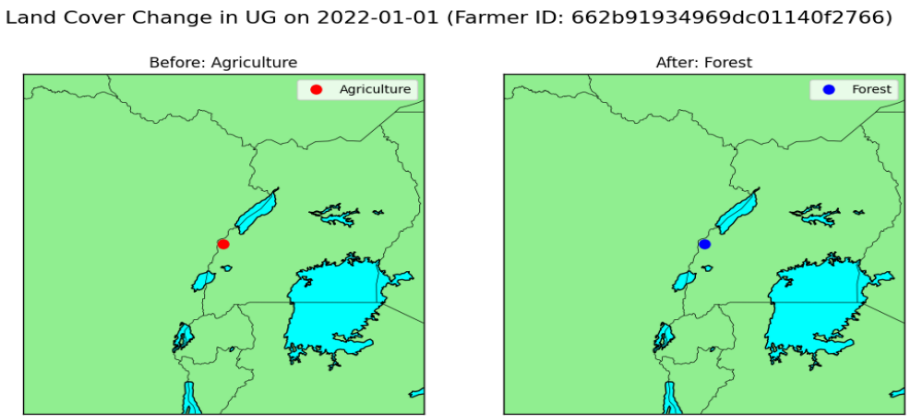

Figure 12: Example of Reforestation in Uganda

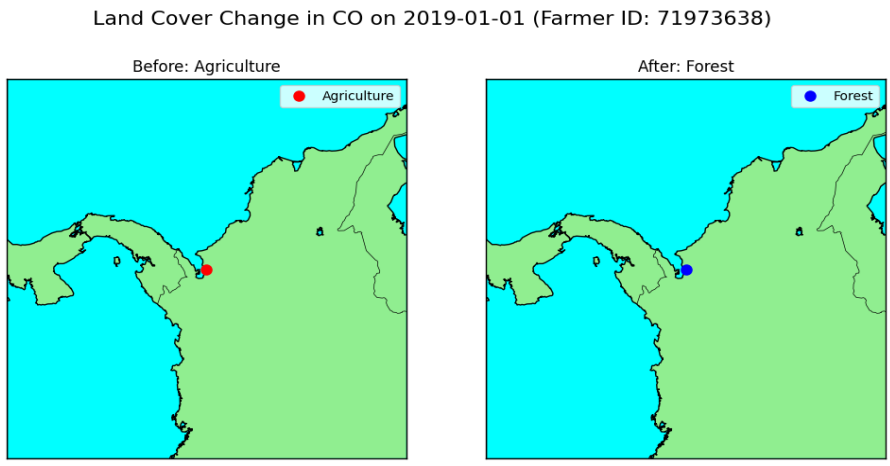

Figure 13: Example of Reforestation in Colombia

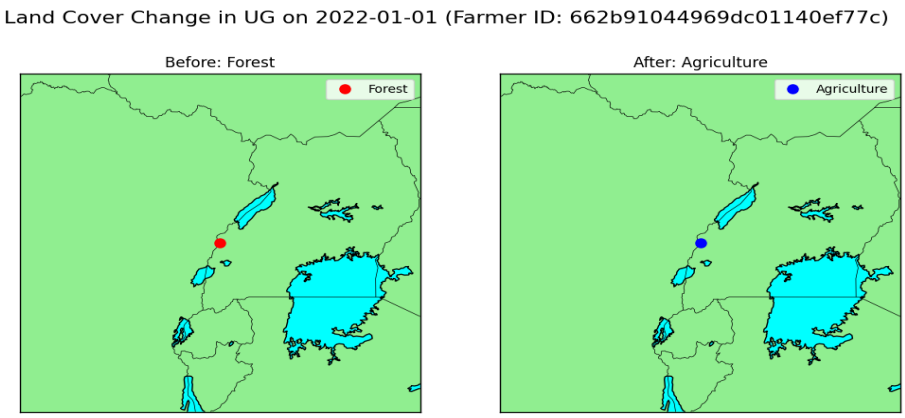

Figure 14: Example of deforestation in Uganda

Observations of Land Cover Changes

The analysis revealed an unexpected trend of widespread reforestation across several regions, contrasting with the global narrative of rampant deforestation for agricultural expansion. Notably, countries like Peru and Uganda, which are significant cocoa producers, showed minimal transitions from agriculture to forest, suggesting that reforestation efforts or natural regrowth are occurring more extensively than previously recognized.

This methodology enables a precise and temporal-specific analysis of land cover changes, which is pivotal for assessing environmental changes and aiding policy decisions. The constraints posed by the dataset timelines necessitate careful handling of data beyond 2022, underscoring the need for updated land cover datasets for future analyses. This approach not only ensures the accuracy of the temporal analysis but also enhances the reliability of the environmental assessments derived from this study.
The observed reforestation across cocoa-producing regions presents both challenges and opportunities. While it may pose short-term economic challenges by reducing land available for cocoa cultivation, it also offers long-term ecological and economic benefits by enhancing ecosystem health and sustainability. This complex interplay of ecological and economic factors should be the focus of continued research to develop strategies that balance environmental sustainability with economic needs.

4.Results and Key Findings

4.1 Production Anomalies

The clustering of Colombian cocoa farmer data was conducted based on geographic coordinates (latitude and longitude), facilitating the identification of spatial patterns and regional differences in farming operations. This geo-based clustering aimed to segment the farmers into groups with similar geographic locations, thereby isolating operational variations that could impact agricultural outputs and sales behaviors. The optimal number of clusters was determined through the silhouette score method, which indicated seven clusters.
The merged dataset, which incorporated both farmer details and sales data with 17841 records, was enhanced with two new columns: 'anomaly' and 'row-anomaly.' These columns flag anomalies at the aggregate level (total sales per farmer) and at the individual transaction level, respectively. The flags (-1 indicating an anomaly) provide a straightforward method for identifying and investigating outlier data points that deviate from expected patterns based on the dataset's historical norms.

Cluster 0

The scatter plot for Cluster 0 illustrates anomalies against the backdrop of total sales and hectares. It is evident that anomalies are scattered across a wide range of hectares, but notably, they appear at the upper extreme values of sales, regardless of the hectare size.

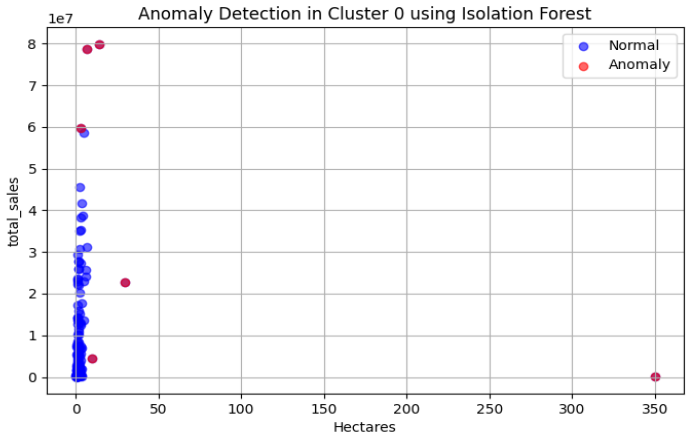

Figure15: Cluster 0 anomalies

Cluster 1

In Cluster 1, anomalies are similarly detected across various hectare sizes but are particularly concentrated at higher sales values. This cluster has anomalies distributed over a range of hectares, indicating that the model picks out both small and large landholdings where sales values do not conform to the typical patterns observed within the cluster.

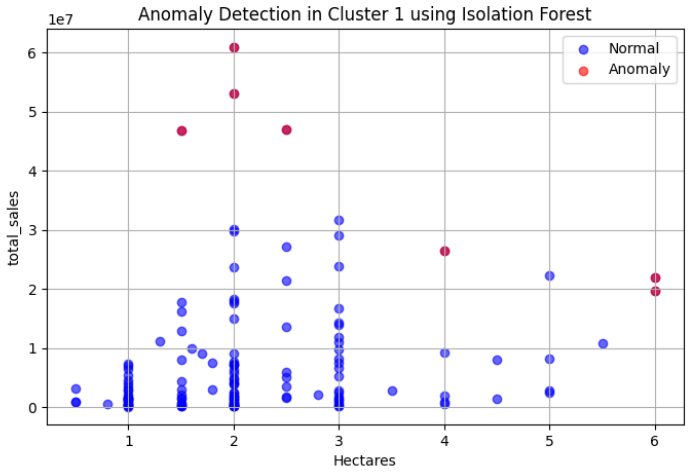

Figure 16: Cluster 1 anomalies

Cluster 2

This cluster's plot shows a dense concentration of normal data points at lower hectares and sales, with anomalies appearing isolated and primarily at higher hectares with significant sales values. The presence of anomalies at these points may indicate exceptional cases where the yield per hectare is exceptionally high or possibly miscategorized sales data.

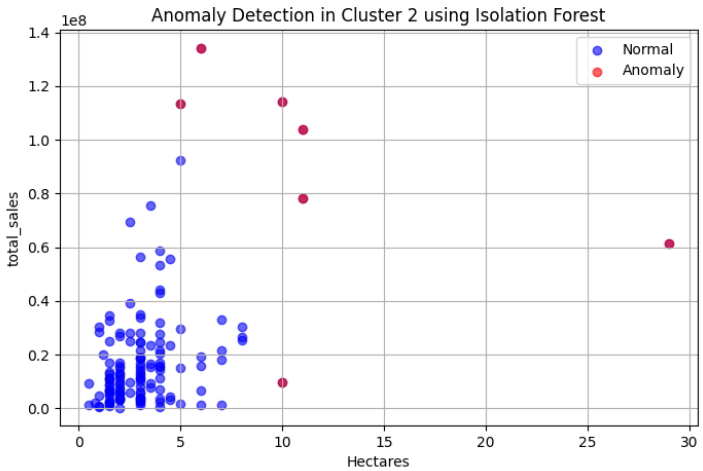

Figure 17: Cluster 2 anomalies

Cluster 3

Anomalies in Cluster 3 are few but are positioned at higher sales values across a broad range of hectares. This indicates a sensitivity of the model to higher revenue figures, which could be due to extraordinary sales achievements or potential discrepancies in sales reporting.

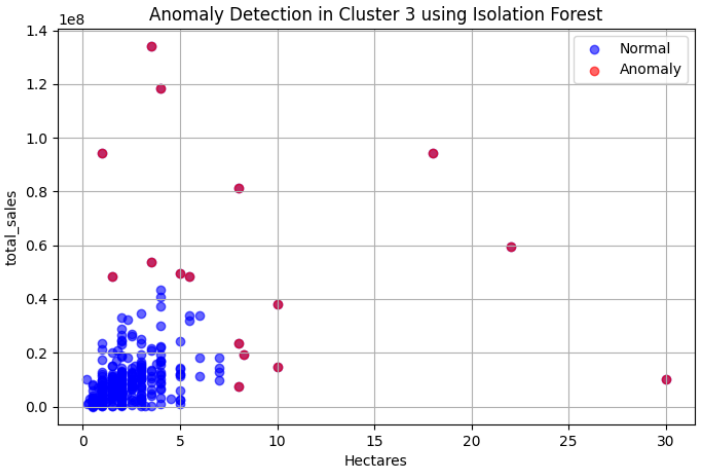

Figure 18: Cluster 3 anomalies

Cluster 4

Cluster 4 reveals anomalies that are very sparse, suggesting that most of the sales data within this cluster falls within expected norms. The few anomalies present span a range of smaller hectare sizes but do not cluster around any particular sales figure, suggesting individual cases of discrepancies.

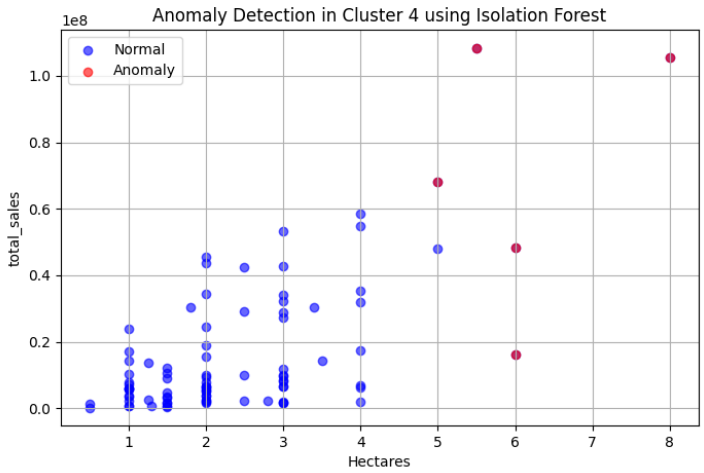

Figure 19: Cluster 4 anomalies

Cluster 5

The anomalies in Cluster 5 are minimal and occur at both ends of the hectare spectrum. This might indicate special cases where sales per hectare are not aligning with the general trends, potentially flagging unique agricultural practices or data entry errors.

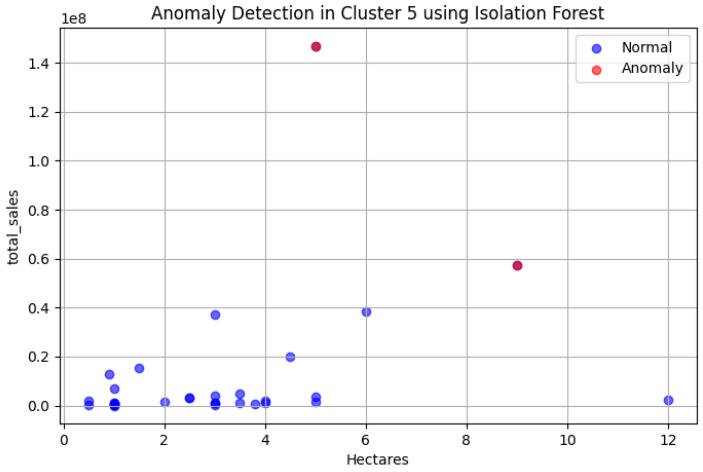

Figure 20: Cluster 5 anomalies

Cluster 6

Finally, Cluster 6 shows a pattern where anomalies are again noted at higher sales figures, spanning a wide range of hectares. Similar to other clusters, these outliers could be indicative of extraordinary cases or discrepancies that merit further qualitative review.

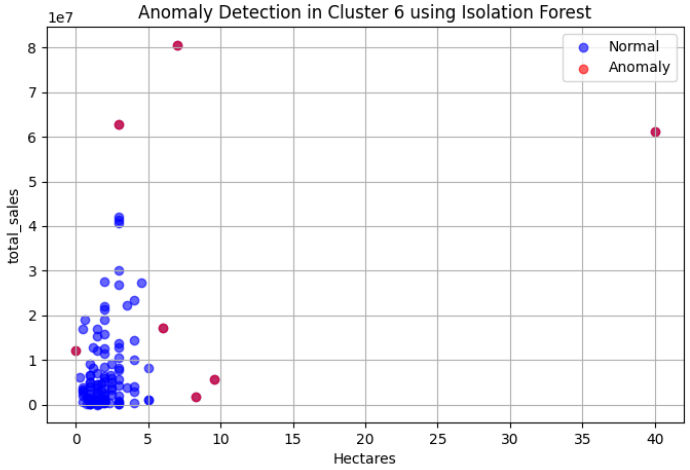

Figure 21: Cluster 6 anomalies

Analyzing the plots from both aggregate and individual transaction levels reveals a consistency in anomaly detection across clusters, as identified by the Isolation Forest algorithm. One notable observation across these plots is the repeated identification of anomalies in large-scale farms, particularly those with extensive hectares under cultivation. For example, in Cluster 0, the farm with 350 hectares is flagged as anomalous in both the aggregated and individual transaction analyses. This consistent flagging suggests potential issues such as data entry errors or misclassification, which might indicate the farmer's declared area being larger than what might be realistically cultivable or managed effectively for cocoa production.

The presence of such anomalies at both levels of analysis underscores the utility of the Isolation Forest algorithm in detecting outliers that may signify underlying data issues or real-world phenomena such as unreported expansion of farmland which can have significant implications for sustainability assessments and compliance checks in agricultural supply chains.

Moreover, the plots also show a range of anomalies across different sizes of land holdings, with smaller land parcels sometimes recording unusually high sales figures, which could indicate high productivity or discrepancies in data reporting. These findings can be crucial for further investigation, potentially leading to on-ground verification of farm sizes and production capacities.

In conclusion, integrating both aggregated and individual transaction-level anomaly detections allows for a more robust understanding of the data, highlighting discrepancies that warrant further investigation to ensure accuracy in reporting and compliance with sustainable farming practices. This dual-level approach enhances the research's contribution to developing methodologies for transparent and reliable data assessment in the cocoa production industry.

The spread and concentration of anomalies in smaller farms across multiple clusters suggest a trend where smaller operations might struggle with accurate reporting or face specific challenges that lead to anomalies in sales data. This could be due to a lack of resources to maintain precise records or complexities in managing smaller-scale productions.

The presence of high-value anomalies in clusters like Cluster 2 could indicate areas where economic incentives to over-report might be higher. Clusters like Cluster 2, showing high-value anomalies, might require deeper investigation to determine the root causes of such discrepancies and to

implement measures that prevent economic exploitation or fraud.
Regions showing frequent anomalies, especially in small farm sizes, might benefit from targeted audits to ensure compliance and accuracy in reporting. Additionally, training programs on record-keeping and data management could help farmers accurately report their production and sales.

4.2 Geospatial Analysis

The land cover classification analysis was conducted by cross-referencing the Copernicus and ESA WorldCover datasets, specifically for the years 2020 and 2021. This allowed for a detailed examination of agricultural land use, deforestation patterns, and possible discrepancies in land classification data for cocoa farming areas across different countries.
The initial classification showed that the majority of the areas reported for cocoa production were correctly categorized as agricultural land. However, the datasets revealed a significant proportion of land classified under other non-agricultural categories, including forest, settlement, bare lands and water bodies. These findings highlight discrepancies between reported farm areas and their corresponding land cover classifications, suggesting possible issues with land use or misreporting.

A total of 1,357 farms were detected in non-harvestable areas (Forest, urban regions, water bodies, or barren land) out of 13,399 data points. These anomalies represent about **10.1%** of the data, suggesting possible misreporting or incorrect land-use classification. The identification of these farms highlights a significant area of concern for cocoa production, particularly as non-harvestable lands should not be classified as agricultural under environmental regulations. This could point to data entry issues, misclassification, or even intentional misreporting.

Additionally, the analysis revealed 59 farms operating outside designated country boundaries, further complicating compliance with international regulations. These cross-border anomalies could result from administrative errors, improper land registration, or intentional misreporting. With the European Union Deforestation Regulation (EUDR) and similar policies requiring strict origin and land-use verification, such discrepancies highlight the need for enhanced geospatial monitoring in the cocoa supply chain.

A significant finding was the presence of 10,710 farms showing cross-dataset discrepancies between the Copernicus and ESA datasets. This means over 80% of the total data points had conflicting land cover classifications between the two data sources. These discrepancies could stem from various factors, including differences in satellite imagery resolution, data processing techniques, or classification methodologies used by the two programs. For instance, one dataset may classify a region as forest while the other might label it as grassland, depending on the resolution or the classification criteria applied. It is important to note that these discrepancies are not solely due to mismatches in classifying non-harvestable areas. The high percentage of discrepancies highlights the need for harmonizing datasets to improve the accuracy of land classification, especially when this data is used for regulatory compliance and sustainability reporting.

The analysis of land cover changes from 2018 to 2022 indicated an unexpected trend toward reforestation rather than deforestation in some regions. This finding diverges from the global narrative that cocoa production often drives deforestation. However, the reforestation observed might be due to several factors, including the effect of regulations such as the EUDR, which may have prompted more sustainable practices or even led to reduced agricultural activity in some areas.
Another potential factor influencing this trend could be the COVID-19 pandemic, which disrupted global supply chains and might have led to a temporary reduction in cocoa production and land use. Farmers may have abandoned or reduced their farming activities due to decreased demand or logistical challenges during the pandemic. Consequently, land classified as agricultural in previous years might have been reclaimed by natural vegetation during this period.

4.3 Economic and Business Perspective

The analysis of cocoa export data from key producer countries Colombia, Uganda, Togo, Ecuador, and Peru reveals notable fluctuations in export volumes between 2018 and 2022. This period saw significant reductions in exports from certain countries, particularly in 2022, which coincided with an observable trend of reforestation in the same regions. The data allows to explore potential links between economic shifts, regulatory pressures such

as the European Union Deforestation Regulation (EUDR), and changes in land use practices.

1.Colombia: Dramatic Decline in 2022

Between 2018 and 2021, Colombia's cocoa exports saw consistent growth, with a total increase of approximately 77%. However, a sharp decline of 47.75% was observed in 2022, marking a dramatic reversal in export trends. Key importers like Estonia (-74.03%), Germany (-38.43%), and Netherlands (-30.77%) significantly reduced their imports. This sudden drop raises questions about the possible impact of stricter regulations such as the EUDR, which could have led to decreased demand for non-compliant cocoa.

From a business perspective, the decline in export volumes poses challenges for the Colombian cocoa industry. Large export markets, particularly in Europe and North America, have reduced their imports, potentially due to more stringent sustainability requirements. This reduction could also lead to economic hardship for farmers who may be pressured to adopt costly sustainability measures or face market exclusion(*The Observatory of Economic Complexity*, 2024).

2. Uganda: Steady Growth Followed by a Major Decline

Uganda experienced strong growth in its cocoa exports between 2018 and 2021, with a 123.7% increase between 2018 and 2019, followed by a steady rise until 2021. However, a significant decline of 44.18% in 2022 mirrored the trend seen in Colombia. Countries like Switzerland (-99.40%), and the United Kingdom (-93.19%) saw drastic reductions in imports.

Uganda's cocoa sector is integral to the livelihoods of many smallholder farmers, and this dramatic fall in exports could have severe economic implications. The decreased demand from major European markets, likely tied to sustainability criteria and compliance with the EUDR, could further exacerbate economic inequalities, especially in rural areas reliant on cocoa farming. There's a need for investment in capacity-building for Ugandan farmers to meet the evolving international standards.

3. Togo: Moderate Decline in 2022

Togo's cocoa exports followed a similar pattern, with steady growth between 2018 and 2021 but an 18.02% decline in 2022. France (-95.34%) and the Estonia (-92.39%) significantly reduced their imports. Unlike the

more dramatic decreases seen in Colombia and Uganda, Togo's decline, while notable, reflects a more moderate shift. It is worth investigating whether Togo's relatively lower decline is due to a less stringent enforcement of sustainable cocoa farming practices or higher compliance levels.

Economically, Togo's cocoa sector remains vulnerable to shifts in international trade policies. The dependence on key markets such as Belgium and the U.S. exposes the country's cocoa farmers to significant risks if these markets continue to adopt more stringent deforestation-related regulations.

4. Ecuador: Resilience Amid Growth

Unlike the previous countries, Ecuador saw a continued increase in cocoa exports, with an 11.67% rise between 2021 and 2022. This growth was achieved despite reductions in imports from major buyers like France (-62.69%) and the Austria (-55.72%). Ecuador's ability to maintain and even grow its export volumes, despite global pressures for sustainability, may point to its relative success in meeting compliance standards, or a shift in trade toward less-regulated markets.

From a business perspective, Ecuador's resilience is notable. It suggests that the country has been able to navigate the challenges posed by sustainability regulations and may have diversified its export markets. However, further analysis is needed to determine whether this growth is sustainable in the long term, particularly as global demand increasingly shifts toward sustainably sourced products.

5. Peru: Gradual Growth with Modest Declines

Peru experienced gradual growth in cocoa exports between 2018 and 2022, with a 3.98% increase in 2022. While countries like Germany (-30.45%) and Spain (-24.56%) reduced imports, the impact was offset by increases from markets such as South Korea (+48.56%) and Canada (+33.78%).

Peru's relatively stable export performance suggests that it has managed to maintain market access despite increasing regulatory pressures. This stability could be attributed to more established sustainability practices or less reliance on European markets. However, the modest decline in European imports indicates that the country, too, will need to continue investing in sustainable farming practices to maintain market share in regions with strict environmental regulations.

The reforestation trends observed across several of these countries, particularly Uganda and Colombia, indicate a potential shift in land use practices. This may be influenced by international pressure to curb deforestation and comply with sustainability standards such as the EUDR. However, the reforestation trend also coincides with significant drops in export volumes, raising concerns about the economic impacts on smallholder farmers.

It is plausible that the declining export figures reflect both a market response to new regulations and a reduction in the availability of agricultural land for cocoa production, as reforestation efforts potentially reduce arable land. Additionally, the data may reflect the aftershocks of the COVID-19 pandemic, which disrupted global trade and could have accelerated these shifts.

5.Limitations and Future work

5.1 Limitations

While the production anomaly detection process has successfully identified outliers and potential discrepancies across various clusters, certain limitations must be considered to contextualize these results.

1. Isolation Forest Sensitivity: The use of the Isolation Forest algorithm effectively detects outliers but is sensitive to parameter settings such as contamination levels and the number of estimators. Adjusting these parameters could lead to variations in the flagged anomalies, potentially affecting the consistency and accuracy of the results. A more detailed hyperparameter tuning process might be required for finer results.

2. Lack of Ground Truth Data for Validation: Without field validation, the flagged anomalies remain theoretical. A significant limitation of the current analysis is the absence of ground truth data on the ground verification that could confirm whether the detected anomalies truly represent discrepancies in production reporting or sales data. Collaboration with certification bodies or field audits would be crucial for verifying the anomalies.

3. External Factors Affecting Production: The detection process doesn't account for external environmental or economic factors that could influence production or sales figures. For instance, unusual weather conditions or sudden market demand could explain high sales figures from small farms, yet these factors are not integrated into the anomaly detection model.

The geospatial analysis in this study, particularly in land cover classification and anomaly detection, presents some challenges and limitations that affect the interpretation of results and the broader applicability of the findings. Below are the key limitations encountered during the study:

1. **Data Accuracy and Resolution**: The accuracy of land cover data varies between datasets, and the spatial resolution of grid cells (300m for Copernicus and 10m for ESA) influences the ability to detect fine-scale changes. The discrepancies between the Copernicus and ESA datasets, where over 80% of data points showed conflicting classifications, illustrate the impact that differences in resolution and classification techniques can have on the results. These discrepancies may lead to misinterpretation of land cover changes, especially in areas where transitions between categories (e.g., forest to agriculture) occur within small, fragmented landscapes. As a result, subtle yet important changes might be overlooked, or false anomalies may be flagged, affecting the reliability of anomaly detection.

2. **Temporal Availability of Data**: The temporal aspect of the data is another limitation. The land cover datasets used in the study cover the years 2018–2022, meaning that any changes occurring after this period remain undocumented. Given the evolving nature of land use, especially in regions undergoing rapid agricultural expansion or reforestation, this lack of recent data limits the ability to draw firm conclusions about ongoing trends. Furthermore, the absence of data from 2023 and 2024 makes it difficult to assess whether observed reforestation trends are continuing or if deforestation has re-emerged as a threat in these regions.

3. **Complex Landscape Dynamics**: Ecological and agricultural landscapes are inherently complex, and standard land cover categories may not adequately capture this complexity. In regions where multiple land use types coexist or where seasonal variations lead to temporary changes in land cover, the classification systems may fail to reflect the full range of

landscape dynamics. For example, a region classified as "forest" in one dataset may be classified as "grassland" in another due to subtle differences in interpretation of vegetative cover. These complexities complicate the anomaly detection process, as not all discrepancies between datasets indicate environmental or regulatory violations.

4. Reforestation Trends and External Factors: While the land cover analysis identified reforestation trends in several regions, it is essential to consider external factors such as the impact of the COVID-19 pandemic, which may have contributed to reduced agricultural activity or the abandonment of farms. These factors may have temporarily affected land use patterns, skewing the data towards reforestation without reflecting long-term trends. Further, the impact of economic shifts and regulatory pressure, such as compliance with the EUDR, may have influenced land use decisions, but these influences are not fully captured due to the temporal limitations of the data.

5. Geospatial Precision and Cross-Border Anomalies: In the detection of cross-border anomalies, geospatial precision plays a critical role. The use of a 1:10 million scale for country boundaries (sourced from Natural Earth) is highly accurate but may still allow for minor discrepancies in the alignment of coordinates, especially along border regions. The detection of 59 cross-border anomalies could be influenced by these small-scale errors in geospatial precision, particularly where borders are contested or imprecisely defined. This limitation needs to be considered when interpreting the implications of cross-border farming activities.

6. Lack of Ground-Truth Verification: The findings from satellite data and cross-dataset comparisons are inherently based on remote sensing techniques. Without ground-truth verification, it is challenging to confirm whether the detected anomalies accurately reflect changes on the ground. For example, areas flagged as non-harvestable or as showing discrepancies between datasets may not necessarily correspond to real-world misreporting or environmental violations. Ground-level data, such as farmer reports or governmental land audits, would provide additional validation but were unavailable for this study.

5.2 Future work

Production Anomalies:

Several avenues for future research could enhance the understanding of anomalies in agricultural data and improve the robustness of anomaly detection methodologies. To deepen the analysis, future studies should consider incorporating more comprehensive datasets, such as weather patterns, market prices, and crop yield data from comparable regions. Cross-referencing detected anomalies with external data can help confirm their validity and reveal underlying causes, such as environmental impacts or market fluctuations.

Additionally, longitudinal studies that track the same data over multiple growing seasons would provide valuable insights into the temporal dynamics of these anomalies. Such studies could help determine whether anomalies are recurring over time, indicating systemic issues, or if they are isolated incidents, potentially due to data entry errors. Future research could also investigate the economic and operational impacts of these anomalies, assessing how they affect farm profitability, sustainability, and efficiency.

Cross-Border Anomalies:

The detection of geospatial anomalies, particularly cross-border farming activities, presents several opportunities for future research. Advanced anomaly detection algorithms, which can operate in real-time, could be developed to automatically detect and flag inconsistencies with geospatial and administrative boundaries. Cross-validation techniques that use multiple data sources, such as satellite imagery, can help validate reported coordinates and reduce the likelihood of erroneous data entries.

Future research should also examine the socio-economic impacts of misreported data, especially regarding its effects on agricultural statistics, policy-making, and resource allocation. Research into the development of policy and regulatory frameworks for standardizing data collection protocols could improve data integrity across countries and regions. Additionally, educational programs for farmers and data collectors can help improve the accuracy of data collection, especially when using GPS devices. Technological improvements in GPS accuracy would further support the goal of precise and reliable data reporting in agricultural

settings.

Land Cover Classification:

Future work in land cover classification can focus on enhancing verification methods and ensuring data accuracy. Stratified random sampling could be implemented to provide a comprehensive understanding of the dataset's accuracy by ensuring all land cover categories are represented in the validation process. Automated tools that cross-reference satellite data with high-resolution imagery from other reliable GIS sources could streamline the verification process, improving efficiency without increasing manual labor.

Continuous validation frameworks should also be established, incorporating regular updates and user feedback to catch new errors and discrepancies. A system that allows users to report inconsistencies in the data would provide an additional layer of accuracy assurance, creating a dynamic and responsive validation process that evolves with the dataset.

Land Cover Changes:

Several hypotheses for future research could be explored to better understand the implications of land cover changes, particularly reforestation, on cocoa production. For instance, future studies could investigate whether reforestation in major cocoa-producing regions correlates with measurable decreases in cocoa exports, impacting local economies. Alternatively, the ecological benefits of reforestation such as improved biodiversity, pollination, and soil health could be quantified to assess how they support long-term agricultural sustainability.

Future studies should also consider the socio-economic adaptations of communities in reforested or deforested areas. Research into how these communities diversify economically or adapt their agricultural practices to maintain or increase output would provide valuable insights into the intersection of environmental conservation and economic viability. A multifaceted research approach, combining economic data analysis, ecological surveys, and socio-economic studies, would provide a holistic understanding of the impacts of land cover changes on cocoa farming regions.

6.Conclusion

This research aimed to develop and implement methodologies for detecting anomalies in agricultural data, particularly focusing on cocoa production in regions like Colombia, Togo, Ecuador, Uganda, and Peru. The primary objective was to enhance transparency, ensure compliance with regulations such as the European Union Deforestation Regulation (EUDR), and support sustainable farming practices. Through the integration of machine learning techniques and geospatial analysis, this thesis provides novel insights into the challenges and opportunities in monitoring and regulating agricultural supply chains.

One of the key contributions of this research is the application of the Isolation Forest algorithm for detecting production anomalies. This methodology allowed for both aggregate-level and individual transaction-level analysis, providing a more comprehensive understanding of discrepancies in sales data relative to land size. The study identified various anomalies in both large- and small-scale farms, suggesting potential issues like data misreporting, unreported farmland expansion, or operational inefficiencies. These findings underscore the importance of anomaly detection in ensuring accuracy in reporting and maintaining the integrity of agricultural supply chains. By flagging outliers, this methodology aids stakeholders in conducting more targeted investigations, which are essential for improving sustainability and transparency in cocoa production.

Geospatial analysis played a critical role in this research, especially in identifying cross-border anomalies and land use discrepancies. By comparing satellite data from the Copernicus and ESA datasets, this study revealed a significant proportion of farms with cross-dataset discrepancies in land cover classification. These discrepancies, which could be attributed to different resolutions and classification methods, highlight the need for harmonizing satellite data to improve the accuracy of land cover assessments. The detection of non-harvestable land classified as agricultural further emphasized the importance of accurate geospatial data in regulatory compliance. Additionally, the identification of cross-border anomalies where farms were reported outside of designated national boundaries points to possible misreporting or errors in data collection, necessitating the implementation of stricter data validation protocols.

Another significant finding was the trend of reforestation observed in the land cover change analysis. While global narratives often emphasize deforestation in cocoa-producing regions, this study found instances of reforestation in countries like Uganda and Colombia. These findings raise important questions about the socio-economic drivers behind these changes, including the potential impact of regulations on farmer livelihoods. The correlation between reforestation and declining cocoa exports, particularly during the COVID-19 pandemic, suggests that factors such as reduced economic activity or changes in land use practices could be influencing these patterns. However, the lack of more recent data for 2023 and 2024 presents a limitation, as a more up-to-date analysis would be necessary to fully understand the long-term effects of these land cover changes on local economies and global cocoa supply chains.

The study also highlights the need for continuous improvement in data collection, validation, and analysis methodologies. The discrepancies in the land cover datasets and the challenges posed by cross-border anomalies underline the importance of developing more sophisticated detection systems that can operate in real-time. Such advancements would allow for quicker identification of data inaccuracies, improving decision-making processes related to agricultural policy, sustainability, and supply chain management. Moreover, integrating more advanced satellite data and enhancing GPS accuracy in agricultural settings would further reduce the occurrence of geospatial discrepancies and improve the overall reliability of the data.

In conclusion, this research contributes to the growing field of agricultural anomaly detection by providing a framework that integrates machine learning, geospatial analysis, and land cover classification. By addressing both production and geospatial anomalies, this study offers a comprehensive approach to improving data transparency and regulatory compliance in agricultural supply chains. The insights gained from this research are particularly relevant for certification bodies, policymakers, and stakeholders aiming to ensure the sustainability of cocoa production while safeguarding the livelihoods of smallholder farmers. Future work in this area should continue to explore the economic and ecological impacts of land use changes, refine anomaly detection algorithms, and leverage technological advancements to create more robust systems for monitoring agricultural activities.

References

An, S. H., Heo, G., & Chang, S. H. (2011, March). *Detection of process anomalies using an improved statistical learning framework*. Elsevier. https://www.sciencedirect.com/science/article/abs/pii/S0957417410006482

Avadi, A. (2023, February 17). *Environmental assessment of the Ecuadorian cocoa value chain with statistics-based LCA*. https://doi.org/10.1007/s11367-023-02142-4

Bass, S., Thornber, K., Markopoulos, M., Roberts, S., & Grieg-Grah, M. (2001). *Certification's impacts on forests, stakeholders and supply chains*. https://books.google.it/books?hl=en&lr=&id=9camrvdw5-sC&oi=fnd&pg=PT5&dq=role+of+certification+and+regilatory+frameworks+on+supply+chains&ots=PujEricA6f&sig=WHETPSfj90SKkj_L9mw0lEhY6Vs&redir_esc=y#v=onepage&q=role%20of%20certification%20and%20regilatory%20frameworks%20on%20supply%20chains&f=false

Bymolt, R., Laven, A., & Tyszler, M. (2018). *Demystifying the Cocoa Sector in Ghana and Côte d'Ivoire*. The Royal Tropical Institute (KIT), Amsterdam, Netherlands. https://www.kit.nl/wp-content/uploads/2020/05/Demystifying-complete-file.pdf

Chen, Y.-M., Li, J.-S., & Chen, T.-Y. (2023). *A Machine Learning-Based Anomaly Detection Method and Blockchain- Based Secure Protection Technology in Collaborative Food Supply Chain*. https://www.igi-global.com/article/a-machine-learning-based-anomaly-detection-method-and-blockchain-based-secure-protection-technology-in-collaborative-food-supply-chain/315789

Chunguang, B., Matthew, Q., & Joseph, S. (2022, June 15). *Analysis of Blockchain's enablers for improving sustainable supply chain transparency in Africa cocoa industry*. https://www.sciencedirect.com/science/article/pii/S0959652622015062

European Commission. (2023, June). *Deforestation Regulation implementation—European Commission*. https://green-business.ec.europa.eu/deforestation-regulation-implementation_en

Ke Zhang, Yi Chai, Simon X. Yang, & Weng, D. (2011, March). *Pre-warning analysis and application in traceability systems for food production supply chains*. https://www.semanticscholar.org/paper/Pre-warning-analysis-and-application-in-systems-for-Zhang-Chai/aa3653a317df4e58c7a2f7b7ba66af6d63e8d6e4

Kwarteng, A. K., & Emefa, A. T. (2023, September 17). *Understanding Sustainable Value Capture for Ghana's Cocoa Farmers on the*

Cocoa-Chocolate Value Chain. https://doi.org/10.5539/jsd.v16n5p145

Lamarche, C., & Defourny, P. (2024, February 22). *Product Quality Assurance Document*. Copernicus.

Manh, B. D., Son, D. H., & Trung, N. L. (2024, July). *Semi-Supervised Learning for Anomaly Detection in Blockchain-based Supply Chains*. https://arxiv.org/abs/2407.15603

Statista. (2023). *Largest cocoa producing countries worldwide 2023/2024*. Statista. https://www.statista.com/statistics/263855/cocoa-bean-production-worldwide-by-region/

The Observatory of Economic Complexity. (2024). https://oec.world/en

Tirkolaee, E., & Sadeghi, S. (2021, June 22). *Application of Machine Learning in Supply Chain Management: A Comprehensive Overview of the Main Areas*. https://doi.org/10.1155/2021/1476043

Trusty. (n.d.). *Trusty Website*. Retrieved 8 September 2024, from https://www.trusty.id/

Van De Kerchove, R. (2020, October 15). *World Cover Product User Manual 2020*. European Space Agency.

Van De Kerchove, R. (2022, October 24). *World Cover Product User Manual 2021*. European Space Agency.

Wainaina, P., A.Minang, P., & Nzyoka, J. (2022, February). *Negative environmental externalities within cocoa, coffee and oil palm value chains in Africa*. https://www.researchgate.net/publication/358404261_Negative_environmental_externalities_within_cocoa_coffee_and_oil_palm_value_chains_in_Africa

Wang, J., & Yue, H. (2017, March). *Food safety pre-warning system based on data mining for a sustainable food supply chain*. https://www.sciencedirect.com/science/article/pii/S0956713516305242