MAKERERE UNIVERSITY

COLLEGE OF ENGINEERING, DESIGN, ART AND TECHNOLOGY

SCHOOL OF BUILT ENVIRONMENT

DEPARTMENT OF GEOMATICS AND LAND MANAGEMENT

AN INTEGRATED GEOSPATIAL PROCESSING PIPELINE FOR EUDR-COMPLIANT COFFEE PLOT GEOLOCATION DATA: A CASE STUDY OF MASAKA DISTRICT

BY

KASAMBULA LEAH

2024/HD08/1517

2400701517

SUPERVISOR: ASSOC.PROF. ANTHONY GIDUDU

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE AWARD OF THE DEGREE OF MASTER OF SCIENCE IN GEO-INFORMATION SCIENCE AND TECHNOLOGY.

JUNE, 2026

DECLARATION

I Kasambula Leah, declare that this thesis is my original work and that, it has never been presented by any person or institution for an academic award.

Signature..... ..... Date: 12th / 06 / 2026.

DEDICATION

First and foremost, I dedicate this thesis to God Almighty for the gift of life, wisdom, strength and guidance throughout my academic journey.

I also dedicate this work to my beloved parents, Mr. and Mrs. Kasambula and to my brother, Shem, whose prayers, unwavering support, encouragement and sacrifices sustained me throughout my studies. Their love and belief in my abilities inspired me to persevere and achieve this milestone.

Finally, I dedicate this work to everyone who believed in me and supported me in one way or another during this journey.

APPROVAL

This thesis has been submitted to the Department of Geomatics and Land Management for review with approval of my university supervisor mentioned below;

Assoc.Prof. Anthony Gidudu



Signature…….. Date: 12/06/2026

ACKNOWLEDGEMENT

I am deeply grateful to Almighty God for His protection, guidance, wisdom and good health throughout my studies and the preparation of this thesis.

My sincere appreciation goes to my supervisor, Assoc.Prof. Anthony Gidudu, for his invaluable guidance and constructive comments throughout this research. His expertise and encouragement greatly contributed to the successful completion of this work.

DIGITISATION AND SELF-ARCHIVING CONSENT AGREEMENT

Digitisation and Self-Archiving Consent Agreement: Theses

Agreement between Makerere University & Students (Authors of Theses / Dissertations / Reports)

1. The author is a student of Makerere University and author of the thesis / dissertation / report titled:

AN INTEGRATED GEOSPATIAL PROCESSING PIPELINE FOR SUBA-COMPLIANT COFFEE PLOT GEOLOCATION DATA : A CASE STUDY OF MUKONO DISTRICT

2. The author grants to the University:
   a. The right to deposit the electronic version of the Thesis / Dissertation / Report into Makerere University Institutional Repositories (Mak IR) or (Mak UD); and
   b. The right to store the thesis / dissertation / report in Mak IR / Mak UD and make it permanently available to the general public via the Internet at no cost to the general public after a grace period (if any is specified). **Choose one of the three options below (c, d, or e):**
   c. The Author may opt for immediate open access to the public .....
   d. Or Restrict access indefinitely: Restrict access indefinitely
   e. Or Restrict access for the specified number of years: .....
   f. Reason for restriction: Further dissemination consideration

3. The author warrants that to the best of the authors knowledge and belief:
   a. The thesis / dissertation / report is an original work;
   b. The author is the owner of all the intellectual property in the thesis / dissertation
      or
   c. The Author is entitled to deal with the intellectual property in the thesis / dissertation / report by publishing it on the Internet
   d. The Author has the right, power and authority to enter into this Agreement and to grant the University the rights contained in this Agreement; and
   e. The University's use of the thesis / dissertation / report pursuant to this Agreement will not infringe the intellectual property rights of any third party.

4. The Author acknowledges and agrees that the University is not responsible or liable for any breach of the intellectual property rights in the thesis / dissertation / report, in particular any breach of copyright, as a result of the use of the thesis / dissertation / report pursuant to this Agreement.

5. The University acknowledges that the rights granted by the Creator in clause 2 of this Agreement, do not cause any transfer or assignment of any proprietary rights in the intellectual property in the article to the University.

Signed by the Author as confirmation that the Author has read and accepted the terms of this Agreement:

Name: Kasambula Leah

College: CEDAT School: School of Built Environment.

(Tick) Type of Degree: (Undergraduate / PGD / Masters ✓ / PhD), Reg. No.: _____

Tel No.: 0786573475 E-Mail: kasambula.leah@students.mak.ac.ug

Signature: _____ Date: 6/11/2026

Supervisor's endorsement: Dr. ANTHONY GIBUNDA 6th/11/2026

MakIRs Policy

TABLE OF CONTENTS

DECLARATION i
DEDICATION ii
APPROVAL iii
ACKNOWLEDGEMENT iv
DIGITISATION AND SELF-ARCHIVING CONSENT AGREEMENT v
LIST OF FIGURES ix
LIST OF TABLES x
ACRONYMS xi
ABSTRACT xii
CHAPTER ONE: INTRODUCTION 1
1.1 Background 1
1.2 Problem Statement 2
1.3 Objectives 3
1.3.1 Main Objective 3
1.3.2 Specific Objectives 3
1.3.3 Research Questions 3
1.4 Significance 3
1.5 Justification 4
1.6 Conceptual Framework 5
1.7 Scope of the study 6
CHAPTER TWO: LITERATURE REVIEW 7
2.1 The Regulatory Framework: EUDR and Global Traceability 7
2.1.1 The Context: Transitioning from Voluntary to Mandatory Governance 7
2.1.2 The Technical Mandate: Spatial Requirements and Standards 7
2.1.3 The Smallholder Equity and Efficacy Dilemma 8
2.2 Technical Barriers to Geospatial Harmonization and Validation 9
2.2.1 Spatial Data Interoperability and Heterogeneity 9
2.2.2 Geometric Integrity and Topological Validation 9
2.2.3 Coordinate Reference System and Geodetic Inconsistencies 10
2.2.4 Toward Integrated and Scalable Solutions 11

2.3 Geospatial Processing Pipelines and Workflow Automation 11
2.3.1 From Tool-Centric Workflows to Pipeline Architectures 11
2.3.2 Orchestration and Workflow Management 12
2.3.3 Reproducibility and Scalability in Geospatial Workflows 12
2.3.4 Standardization, Reproducibility and Emerging Practices 13
2.3.5 Implications for Geospatial Compliance Workflows 13
CHAPTER THREE: METHODOLOGY 15
3.1 Research Design 15
3.2 Study Area 15
3.3 Data Sources and Characteristics 16
3.3.1 Secondary Data 16
3.3.2 Reference Data 17
3.4 Methodology Overview 17
3.5 Phase I (Specific Objective 1): Data Characterization 19
3.5.1 Data Extraction, Ingestion and Master GeoDataFrame Construction 19
3.5.2 Geometry Classification, CRS Inventory, Area Analysis 20
3.5.3 Spatial Data Quality Assessment 20
3.5.4 Descriptive Frequency Analysis 21
Table 6: Descriptive Frequency Analysis 21
3.6 Phase 2 (Specific Objective 2): Pipeline Design, Implementation and Design 22
3.6.1 Functional Requirement Definition 22
3.6.2 Bridge File Loading, data Pipeline Initialization and work flow 24
3.6.4 EUDR-IS Compliance Decision Gate and Output Generation 27
3.6.5 Performance Evaluation 28
CHAPTER FOUR: RESULTS AND DISCUSSION 30
4.1 Overview of the characterization results (Specific objective 1) 30
4.1.1 File format distribution 32
4.1.2 CRS distribution 32
4.1.3 Geometry validity and topological integrity assessment 32
4.1.4 Attribute schema-Area consistency check 33
4.2 Design, Implementation and Performance Evaluation of the Integrated Geospatial Processing Pipeline (Specific Objective 2) 34
4.2.1 Pipeline Architecture Overview 34

4.2.2 Processing Throughput 34
4.2.3 Automated Transformation Reliability (ATR) 35
4.2.4 Compliance Yield 35
CHAPTER FIVE: CONCLUSION AND RECOMMENDATIONS 36
5.1 Conclusion 36
5.1.1 Conclusion for Specific Objective 1: Data Characterization 36
5.1.2 Conclusion for Specific Objective 2: Pipeline Development and Evaluation 36
5.2 Recommendations 37
1) Implement Upstream Field Validation 37
2) Adopt Automated Geospatial Compliance Processing Workflows 37
3) Develop a QGIS-Based Implementation of the Pipeline 37
4) Future Research 37
REFERENCES 38
APPENDICES 41
Appendix A: Runtime Execution Sequence Diagram 41
Appendix B: Sample Audit Report 42
Appendix C: Sample CRS Inventory Result 42
Appendix D: Sample Area Consistency check 42
Appendix E; Process Throughput 43
Appendix F: Processing log 43

LIST OF FIGURES

Figure 1: Theoretical and conceptual framework for automated EUDR geolocation data preparation 5

Figure 2: Location of the study area (Masaka District) 16

Figure 3:Illustrates the overall methodological workflow adopted in the study. 18

Figure 4: Masaka Coffee Plot Dataset Characterization findings..... 31

LIST OF TABLES

Table 1: Application of Design Science Research (DSR) Stages in the Study 15

Table 2: Reference Data..... 17

Table 3: Back-end logic: Data Extraction, Ingestion and Master GeoDataFrame Construction.. 19

Table 4: Back-end logic: Geometry Classification, CRS Inventory and Area analysis 20

Table 5: Back-end Logic: Spatial data quality assessment..... 20

Table 6: Descriptive Frequency Analysis..... 21

Table 7: Derived Functional Requirements 22

Table 8: Back-end Logic Module 1: Format Detection & Conversion 24

Table 9: Back-end Logic Module 2: CRS Identification and Reprojection..... 25

Table 10: Back-end Logic Module 3: Geometry Validation & Repair 25

Table 11: Back-end Logic Module 4: Attribute Schema Enforcement 26

Table 12: Pipeline Module Architecture and Primary Functions 34

ACRONYMS

EU	European Union
EUDR	European Union Deforestation Regulation
EUDR IS	European Union Deforestation Regulation Information System
WGS	World Geodetic System
GIS	Geographical Information System
CRS	Coordinate Reference System
PII	Personally Identifiable Information
UBOS	Uganda Bureau of Statistics
os	Operating System
WfMS	Scientific Workflow Management Systems
GUI	Graphical User Interface
OGC	Open Geospatial Consortium
KML	Keyhole Markup Language
ERR	Error
ATTR	Attribute

ABSTRACT

The European Union Deforestation Regulation (EUDR) requires operators placing coffee and other regulated commodities on the European market to submit geolocation data for production plots as part of their due diligence obligations. Although the regulation is format-neutral, the EUDR Information System is intended to support standardized geolocation submissions, typically aligned with GeoJSON structures in the World Geodetic System 1984 (WGS84) coordinate reference system. In Uganda's coffee sector, geolocation data are collected by multiple actors using heterogeneous tools and workflows, resulting in inconsistencies in file formats, coordinate reference systems, geometry quality and attribute structures, thereby complicating compliance preparation. This study developed a geospatial processing pipeline to transform heterogeneous coffee farm geolocation datasets from Masaka District into EUDR submission-ready GeoJSON outputs. The study first characterized technical inconsistencies within 1,350 aggregated coffee farm geolocation datasets comprising 1,252 polygons and 98 point features. The datasets were found to exist in five different file formats and two coordinate reference systems, with additional inconsistencies relating to geometry validity, topology, area reporting and coordinate precision. Based on these findings, an integrated geospatial processing pipeline was implemented in Google Colab to automate dataset standardization, validation and compliance preparation. Evaluation of the developed pipeline demonstrated an overall processing throughput of 78.49 features per second and generated 1,202 submission-ready GeoJSON outputs, 147 flagged outputs and one rejected output from the assessed datasets. The results indicate that the pipeline successfully transformed heterogeneous geolocation datasets into EUDR submission GeoJSON outputs while maintaining strong operational reliability and minimizing manual intervention. The study demonstrates that integrated geospatial processing workflows can provide an efficient, reliable and reproducible approach for preparing coffee farm geolocation data for regulatory compliance. The developed pipeline offers a practical solution for coffee exporters operating within fragmented supply chains and contributes to the broader application of automated geospatial data standardization for environmental governance and agricultural commodity traceability.

Keywords: *EUDR, GeoJSON, geospatial pipeline, coffee geolocation, GIS automation*

CHAPTER ONE: INTRODUCTION

1.1 Background

Global concern over commodity-driven deforestation has increasingly shaped environmental governance and international agricultural trade. Agricultural expansion linked to globally traded commodities such as; coffee, cocoa, soy and palm oil has been widely identified as a major driver of forest loss and the associated biodiversity decline (Kristin et al., 2022). In response, policy frameworks are increasingly emphasizing accountability within global commodity supply chains, linking continued market access to demonstrable environmental compliance and improved transparency. One of the most prominent of such frameworks is Regulation (EU) 2023/1115, commonly referred to as the European Union Deforestation Regulation (EUDR), which requires that commodities entering the European market be verified as not associated with deforestation occurring after 2020 (European Commission, 2023).

The EUDR introduces stringent traceability requirements by mandating the submission of farm-level geolocation data as part of due diligence processes. For each production site, operators are required to provide precise geographic coordinates and where farm units exceed four hectares, spatial boundaries must be represented as polygons. These datasets are submitted through the EUDR Information System (EUDR-IS), which enforces technical specifications such as the World Geodetic System 1984, accurate geometry and standardized digital formats to support consistent spatial verification (Simonetti et al., 2025). As a result, accurate and standardized geospatial data management is becoming central to regulatory compliance for commodity exporters.

These requirements carry significant implications for Uganda, where the coffee sector is strongly oriented towards European markets. The European Union accounts for approximately 61% of Uganda's coffee exports, making it the country's largest export destination (UCDA, 2025). Coffee production also plays a critical role in the national economy, contributing over 3% to GDP and generating between 20–30% of foreign exchange earnings while supporting millions of smallholder farmers earnings (Clement & Joel, 2022). Given this economic dependence, sustained access to European markets will increasingly depend on the ability of actors across the coffee value chain to meet EUDR traceability and geolocation requirements.

However, implementing these requirements within Uganda's coffee sector presents substantial technical challenges. Coffee production is dominated by approximately 1.7–1.8 million smallholder farmers cultivating small and spatially fragmented plots, often embedded within complex agroforestry systems (International Labour Organization, 2024). Correspondingly, geolocation data are collected and aggregated by multiple actors including; traders, cooperatives and exporters while using diverse tools and field mapping approaches. This decentralized process results in large volumes of geospatial data characterized by inconsistencies in coordinate reference systems, file formats, geometry validity and attribute structures.

Such variability limits the direct usability of these datasets within the EUDR Information System, which requires standardized coordinate systems, formats such as GeoJSON and geometrically valid spatial features. Consequently, the need for systematic approaches to harmonize, validate and standardize geospatial datasets is becoming increasingly critical for ensuring compliance with EUDR requirements.

1.2 Problem Statement

The European Union Deforestation Regulation requires operators placing coffee on the EU market to submit geolocation data for production plots as part of their due diligence obligations (European Commission, 2023). Although, the regulation is format neutral, the submission portal (EUDR-IS) is expected to support standardized geolocation data, typically aligned with GeoJSON structure in the World Geodetic System 1984 (WGS84) coordinate reference system (Simonetti et al., 2025).

In fragmented sourcing contexts such as Masaka District (UCDA, 2020), coffee exporters aggregate geolocation data from numerous smallholder farmers, resulting in large collections of heterogeneous spatial datasets. These datasets commonly differ in file formats, coordinate reference systems, geometry validity and attribute structures. Therefore, they require preprocessing operations such as format conversion, coordinate reference system transformation, and geometry validation prior to submission.

While existing GIS software provides tools for these operations, they are typically applied at tool level rather than within an integrated workflow. As a result, processing is often sequential and uncoordinated, leading to inefficiencies, increased processing time, potential inconsistencies in output datasets and risks of submission rejection under regulatory validation checks.

Currently, there is a lack of a standardized, automated and reproducible geospatial processing pipeline specifically designed to transform high volume heterogeneous geolocation datasets into EUDR- submission ready GeoJSON outputs within a unified workflow. This gap presents a significant technical and operational challenge for exporters managing complex spatial datasets under strict regulatory requirements.

1.3 Objectives

1.3.1 Main Objective

To develop an integrated geospatial data processing pipeline that prepares Masaka District coffee farm geolocation datasets for submission to the EUDR Information System.

1.3.2 Specific Objectives

i. To characterize the technical inconsistencies in the aggregated coffee farm geolocation datasets from Masaka District relative to the EUDR-IS spatial requirements.
ii. To designed, implement and evaluate an integrated geospatial processing pipeline for transforming Masaka coffee farm geolocation datasets into EUDR submission ready GeoJSON outputs.

1.3.3 Research Questions

i. What are the technical gaps between aggregated Masaka District coffee geolocation data and EUDR Information System spatial requirements?

ii. How can an integrated geospatial processing pipeline be designed, implemented and evaluated to transform Masaka coffee farm geolocation datasets into EUDR submission ready GeoJSON outputs?

1.4 Significance

This study is significant as it will provide a practical solution for the coffee exporters in Masaka District to efficiently prepare geolocation data for submission to the EUDR-IS. By integrating the automation of CRS transformation, geometry validation and format conversion into a single automated pipeline, the study will not only provide means to minimize errors but also enhances compliance with EUDR geolocation requirements. Beyond its practical benefits, the study will academically contribute to how geospatial data standardization and topological validation can be

combined into a reproducible workflow hence offering insights into error propagation, CRS transformation effects and geometry correction in highly heterogenous datasets. Finally, from a socio-economic perspective, the pipeline will facilitate smoother access to the EU market by improving the competitiveness and sustainability of coffee exporters operating within fragmented supply chains.

1.5 Justification

With EUDR compliance enforcement scheduled to begin on 30 December, 2026 for large and medium operators and on 30 June, 2027 for micro and small enterprises (European Council, 2025), actors along Uganda's coffee supply chain must act promptly so as to maintain access to the EU market. However, the highly fragmented and multi-actor nature of these supply chains results in heterogeneous geospatial datasets, which complicate validation and increase the time and effort required for compliance preparation.

This need is further reinforced by Uganda's classification as a standard risk country under the EUDR benchmarking system, which requires operators to undertake full due diligence procedures for regulated commodities.

Therefore, this study is justified as it seeks to develop an integrated geospatial processing pipeline capable of detecting and correcting inconsistencies to ensure operational efficiency, reproducibility and the production of EUDR-ready geolocation datasets for submission workflows.

1.6 Conceptual Framework

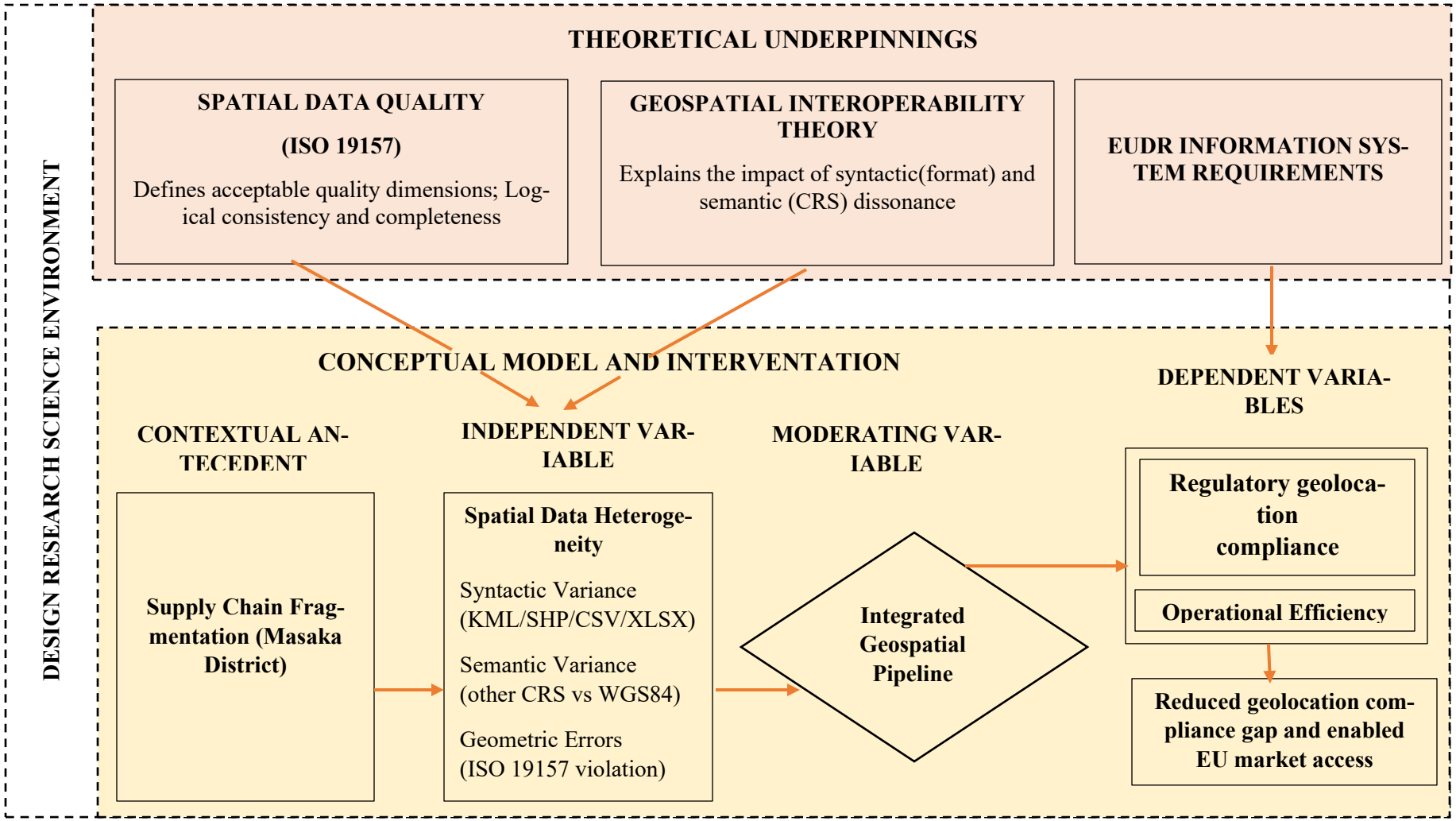


Figure 1: Theoretical and conceptual framework for automated EUDR geolocation data preparation

1.7 Scope of the study

This study was conducted using aggregated coffee farm geolocation datasets obtained from Masaka District, Uganda. The scope focused on the characterization of technical inconsistencies within the datasets and the development of an integrated geospatial processing pipeline for preparing geolocation data to meet the spatial requirements of the EUDR Information System (EUDR-IS). The characterization component was limited to the assessment of file formats, coordinate reference systems, geometry validity, topology integrity, area consistency and coordinate precision. Based on the identified inconsistencies, an integrated processing pipeline was developed to automate file format standardization, coordinate reference system standardization, geometry validation and repair, attribute schema validation and GeoJSON generation. Pipeline evaluation was limited to processing throughput, transformation reliability and compliance yield metrics. The study utilized coffee farm geolocation datasets represented as point and polygon geometries originating from multiple source formats including CSV, XLSX, Shapefile, GeoPackage and GeoJSON.The study did not include deforestation screening, forest-cover change assessment, EUDR risk assessment, direct integration with the EUDR Information System or broader supply chain traceability processes. Therefore, the developed pipeline focused exclusively on preparing geolocation datasets for EUDR-IS submission rather than determining full EUDR compliance.

CHAPTER TWO: LITERATURE REVIEW

2.1 The Regulatory Framework: EUDR and Global Traceability

2.1.1 The Context: Transitioning from Voluntary to Mandatory Governance

For a number of decades, global efforts to mitigate commodity-driven deforestation have relied predominantly on Voluntary Sustainability Standards (VSS) and private corporate commitments (Lambin & Furumo, 2026; Wardell et al., 2021). While these mechanisms have played an important role in piloting cooperative governance approaches, their effectiveness has been constrained by uneven adoption, weak enforcement mechanisms and limited accountability across complex supply chains (Durán & Scott, 2022). As a result, their capacity to deliver large-scale and verifiable forest protection outcomes has remained limited.

The enactment of Regulation (EU) 2023/1115, commonly referred to as the European Union Deforestation Regulation, represents a fundamental shift from voluntary compliance toward a mandatory, enforcement-driven regime (Kleinschmit et al., 2023) . Under this framework, market access to the European Union is contingent upon the submission of verifiable due diligence statements, supported by spatially explicit evidence of production origin (European Commission, 2023). This transition reconfigures governance from a predominantly declarative model to one grounded in measurable and auditable geospatial data. However, this shift also introduces new technical and operational burdens, particularly for actors lacking advanced geospatial capabilities.

2.1.2 The Technical Mandate: Spatial Requirements and Standards

A defining feature of the EUDR is its reliance on geospatial data to enable automated compliance verification through the EUDR Information System. Central to this requirement is the submission of farm-level geolocation data in GeoJSON format, referenced to the World Geodetic System 1984 (WGS84) coordinate reference system (Simonetti et al., 2025). This standardization is intended to ensure interoperability with the European Commission's Global Forest Cover baseline (GFC2020), which serves as the reference dataset for assessing deforestation against the 31 December 2020 cutoff date.

While the adoption of a uniform data standard facilitates large-scale automated screening, it simultaneously assumes that upstream datasets are both geometrically valid and correctly projected. In practice, this assumption is frequently violated, particularly in smallholder contexts where

spatial data is collected using heterogeneous devices, varying coordinate systems and inconsistent levels of technical expertise. Errors arising from improper coordinate reference system (CRS) transformations, including datum misalignment and precision loss, can lead to significant positional inaccuracies that compromise compliance assessments.

In addition, the regulation introduces a scale-dependent requirement whereby production units exceeding four hectares must be represented as polygons rather than points (European Commission, 2023; Simonetti et al., 2025). While this improves the spatial fidelity of larger landholdings, it also introduces challenges related to geometry construction and validation. Common issues such as self-intersections, unclosed rings and topological inconsistencies can render datasets invalid for submission, yet these errors are rarely addressed systematically within existing workflows.

The reliance on standardized formats and geometries enables platforms such as EUDR-IS to perform automated checks against deforestation layers. However, this automation effectively shifts the burden of data quality assurance upstream, requiring producers and intermediaries to ensure that submitted datasets meet strict structural and spatial validity criteria prior to ingestion.

2.1.3 The Smallholder Equity and Efficacy Dilemma

Despite its technical sophistication, the EUDR exposes a significant mismatch between regulatory expectations and on-the-ground realities, particularly in smallholder-dominated systems. Studies of coffee supply chains in countries such as Ethiopia and Uganda point to a clear compliance capacity gap. While awareness of the regulation is growing, many actors lack the technical means to produce valid and standardized geospatial data (Berihun, 2026).

This challenge is compounded by the nature of smallholder agriculture itself. Production often takes place on small, fragmented plots embedded within complex agroforestry systems, making data collection and validation inherently difficult. The effort required to clean, harmonize, and prepare such data introduces additional costs, creating what has been described as a “distributional risk” within supply chains (Afriyie-Kraft et al., 2026). In practical terms, this raise concerns that smallholders may be excluded from European markets in favor of larger producers whose operations are easier to map and verify (Berihun, 2026; Lambin & Furumo, 2026). A further limitation lies in how existing geospatial processes are typically implemented. Tasks such as CRS transformation, geometry validation and format conversion are often handled separately, using different

tools and workflows. This fragmented approach not only reduces efficiency but also increases the likelihood of errors, particularly when dealing with large and heterogeneous datasets.

Taken together, these challenges point to a clear gap in both research and practice. There is a need for integrated, automated geospatial processing pipelines that can reliably transform raw field data into EUDR-compliant outputs. Addressing this gap is critical, not only for improving technical efficiency, but also for ensuring that the shift toward data-driven regulation does not inadvertently disadvantage the very producers it seeks to regulate.

2.2 Technical Barriers to Geospatial Harmonization and Validation

2.2.1 Spatial Data Interoperability and Heterogeneity

The effectiveness of any automated geospatial pipeline is closely tied to the interoperability of its input data. In practice, however, spatial data within the Ugandan context is often fragmented across institutions, reflecting both organizational and technical limitations in the national spatial data infrastructure (SDI) (Musinguzi et al., 2012). Different actors frequently adopt their own naming conventions, coding schemes and classification systems for representing the same geographic features, resulting in a form of semantic heterogeneity that complicates data integration (Du, 2026). This lack of harmonization creates a data silo effect, where datasets cannot be easily combined without substantial preprocessing. More importantly, interoperability challenges at this stage tend to propagate downstream: inconsistencies in structure and meaning often translate into geometric and positional errors once datasets are merged. As such, interoperability is not an isolated concern, but a foundational issue that shapes the reliability of all subsequent geospatial processing steps.

2.2.2 Geometric Integrity and Topological Validation

Beyond interoperability, the geometric integrity of spatial data is a critical requirement for regulatory compliance. Under the EUDR, the validity of a plot is determined not only by its location but also by the structural correctness of its geometry. However, primary data collected in smallholder environments is frequently affected by errors introduced during field mapping and digitization (Verma, 2024).

Geospatial data quality is typically assessed along three key dimensions: positional accuracy, logical consistency (topology) and completeness (Biljecki et al., 2023). In practice, these dimensions are often compromised when using mobile mapping technologies under field conditions. Common issues include; unclosed polygon rings, self-intersections, overlapping features and the creation of

sliver polygons and these are collectively referred to as topological noise (Othman & Idris, 2023). While such errors may appear minor, they have significant implications in automated systems where strict validation rules are enforced.

Datasets that violate these rules risk rejection during automated validation within the EUDR Information System. Consequently, geometric validation cannot be treated as an optional quality check, it must be an integral and automated component of any processing workflow. However, existing practices often rely on manual or semi-automated corrections, which are difficult to scale and prone to inconsistency, especially when dealing with large and heterogeneous datasets

2.2.3 Coordinate Reference System and Geodetic Inconsistencies

A further layer of complexity arises from inconsistencies in coordinate reference systems, particularly in contexts where legacy geodetic frameworks remain in use. In Uganda, a significant portion of spatial data has historically been captured using the Arc 1960 datum (Musinguzi et al., 2012), which is not directly compatible with the WGS84 system required for EUDR-IS submissions. Transforming between these systems is not always straightforward and if done incorrectly, can introduce substantial positional errors.

Empirical evidence highlights recurring issues associated with CRS handling in equatorial regions. These include incorrect hemisphere assignments and failures to account for false northing adjustments when crossing the equator. Such errors can result in what is often described as spatial drift, where mapped features are displaced by tens or even hundreds of meters relative to their true positions (Brazeal et al., 2016). In the context of digitizing farm boundaries depending on alignment with high-resolution satellite imagery, these discrepancies can lead to incorrect classification outcomes.

Importantly, CRS inconsistencies do not operate in isolation. Positional errors introduced during transformation can compromise geometric validity, for example by distorting polygon boundaries or creating unintended overlaps. This highlights the interdependent nature of geospatial data quality, where issues in one processing stage can cascade into others. Despite this, many existing workflows treat CRS transformation as a discrete step, rather than integrating it within a broader validation framework.

2.2.4 Toward Integrated and Scalable Solutions

Taken together, these challenges including; interoperability constraints, geometric and topological errors, CRS inconsistencies and spectral ambiguity highlight the fragmented nature of current geospatial data processing practices. Although each issue has been addressed in isolation within existing studies, there is limited evidence of approaches that treat them as interconnected components of a single workflow.

In practice, these barriers are deeply interdependent. Semantic inconsistencies in source data can lead to errors during integration, CRS misalignments can distort geometries and geometric inaccuracies can, in turn, affect the reliability of downstream validation processes. Addressing these challenges in a piecemeal manner not only reduces efficiency but also increases the risk of cumulative error propagation.

This points to a critical gap in both research and application: the absence of integrated, automated geospatial pipelines capable of handling these interrelated processes in a consistent and reproducible manner. Such pipelines are essential for transforming heterogeneous, error-prone field data into standardized, EUDR-compliant outputs at scale. Without them, the technical burden of compliance remains disproportionately high, particularly for smallholder-dominated systems where data complexity is greatest.

2.3 Geospatial Processing Pipelines and Workflow Automation

2.3.1 From Tool-Centric Workflows to Pipeline Architectures

Geospatial analysis is increasingly moving away from traditional tool-centric workflows toward more integrated and automated pipeline architectures (Radosevic, 2023). Historically, most GIS-based analyses have been carried out through the manual chaining of discrete operations within graphical user interfaces, such as ArcGIS Model Builder or the QGIS Graphical Modeler (Hu et al., 2025; Radosevic, 2023). While these environments have made geospatial tools more accessible, particularly for non-programmers, they often obscure the underlying logic of data transformations. As a result, workflows tend to function as black boxes, where critical processing steps are difficult to trace, reproduce or adapt.

This mode of working also encourages a focus on low-level operational details such as tool parameters and file management rather than on the broader analytical logic of the workflow. In contexts that require the handling of complex and heterogeneous datasets, such as those associated

with regulatory compliance, this lack of abstraction becomes a significant limitation. It restricts the ability to design workflows that are not only functional, but also transparent, reusable, and adaptable across different datasets and use cases.

2.3.2 Orchestration and Workflow Management

A central limitation of traditional GIS workflows lies in their weak support for orchestration. Processes are typically executed in a linear and loosely coordinated manner, with limited capacity to manage dependencies between tasks or to handle more complex, non-linear processing structures (Mansourian & Oucheikh, 2026). This becomes particularly problematic in modern geospatial applications, where data processing often involves multiple interdependent steps, conditional logic and iterative operations.

Scientific Workflow Management Systems, such as KNIME, Pegasus, and Kepler, have emerged to address these challenges by introducing more structured approaches to workflow design and execution (Alam et al., 2023). These systems allow workflows to be modularized into smaller, reusable components that can be combined into larger pipelines, supporting both flexibility and scalability. The ability to nest sub-workflows and explicitly define task dependencies enables a level of control that is largely absent in traditional GUI-based approaches.

However, despite these advances, the adoption of such systems in applied geospatial contexts remains uneven. In many cases, workflows are still constructed in an ad hoc manner, limiting their potential for reuse and integration. This gap is particularly evident in domains where multiple preprocessing steps such as data harmonization, validation and transformation must be coordinated within a single, coherent process.

2.3.3 Reproducibility and Scalability in Geospatial Workflows

The limitations of tool-centric approaches are further reflected in ongoing concerns about reproducibility within geospatial research. Because many traditional workflows rely on manual execution, intermediate steps and parameter choices are often poorly documented, making it difficult to replicate results reliably (Schlögl et al., 2026). This issue is not merely academic, in regulatory contexts, the inability to reproduce processing steps can undermine the credibility and auditability of submitted data.

Scalability presents an additional challenge. Conventional Extract, Transform, Load (ETL) processes are often designed to operate on single machines and struggle to accommodate the volume, variety, and velocity associated with contemporary geospatial datasets (Al-Yadumi et al., 2021; Masson et al., 2022). This is particularly relevant in applications involving large numbers of spatial features or high-resolution datasets, where performance constraints can quickly become prohibitive.

Automated pipeline architectures offer a way forward by enabling parallel processing and the integration of domain-specific libraries tailored to geospatial data. Frameworks such as TorchGeo and Eo-learn, for example, are designed to handle complexities such as multi-resolution imagery and diverse coordinate systems while maintaining computational efficiency (Stewart et al., 2025). By embedding these capabilities within structured workflows, pipelines can achieve both scalability and consistency, reducing the need for manual intervention.

2.3.4 Standardization, Reproducibility and Emerging Practices

Alongside automation, there is a growing emphasis on standardization as a means of improving the transparency and reusability of geospatial workflows. The adoption of open workflow specifications, such as the Common Workflow Language (CWL), enables the formal definition of processing steps in a way that is both platform-independent and easily shareable (Khan et al., 2019). This aligns with the broader push toward FAIR (Findable, Accessible, Interoperable, and Reusable) principles in scientific research (Wilkinson et al., 2016).

In parallel, technologies such as containerization are increasingly being used to encapsulate software environments, ensuring that workflows can be executed consistently across different systems (Fortin et al., 2025; Radosevic, 2023). These developments allow not only for the automation of processing steps, but also for the automatic capture of provenance information, creating a transparent record of how data has been transformed. In effect, this establishes a “single source of truth” that links raw data inputs to final analytical outputs.

2.3.5 Implications for Geospatial Compliance Workflows

Taken together, these developments highlight a broader transition toward more structured, automated and reproducible approaches to geospatial data processing. However, a gap remains

between the capabilities offered by modern pipeline architectures and their application in domain-specific contexts such as regulatory compliance.

In particular, existing implementations often focus on general-purpose data processing or large-scale environmental modeling, with limited attention to the specific challenges associated with heterogeneous, smallholder-derived datasets. Tasks such as coordinate system harmonization, geometric validation and format standardization which are central to compliance frameworks like the EUDR are frequently handled as separate processes rather than as components of an integrated pipeline.

This fragmentation mirrors the challenges identified in earlier sections and reinforces the need for workflows that explicitly account for the interdependencies between different processing stages. An effective geospatial compliance pipeline must therefore go beyond automation in a narrow sense and instead provide a cohesive framework capable of handling data heterogeneity, ensuring geometric integrity and producing standardized outputs in a reproducible manner. Bridging this gap is essential for translating advances in workflow automation into practical solutions for real-world regulatory requirements.

CHAPTER THREE: METHODOLOGY

This chapter describes the systematic approach and technical framework which was employed to develop an integrated geospatial processing pipeline for EUDR submission ready coffee plot geolocation data. The methodology was structured to transition from attaining data, data anonymization, initial data characterization to the design, implementation and evaluation of an automated pipeline.

3.1 Research Design

This study adopted the Design Science Research (DRS) methodology, as formalized by (Hevner et al., 2004) . DRS was selected because the primary research output is an artifact, the integrated geospatial processing pipeline whose value will be demonstrated through evaluation against operationalized performance criteria derived from the real-world problem context.

The DSR approach involved:

Table 1: Application of Design Science Research (DSR) Stages in the Study

DSR Stage	Application in this Study
Problem Identification	Characterization of technical inconsistencies in aggregated coffee farm geolocation datasets
Objective Definition	Derivation of functional requirements from characterization findings
Design and Development	Development of the integrated geospatial processing pipeline in Google Colab
Demonstration	Execution of the pipeline using coffee farm geolocation datasets from Masaka District
Evaluation	Assessment of throughput, transformation reliability and compliance yield

3.2 Study Area

The geographic focus was Masaka District, which is located in the south-central region of Uganda. The district is situated within Uganda's primary Robusta coffee growing belt and was selected as the study area on the basis of three criteria:

1) Its status as one of the major Robusta coffee-producing districts in Uganda.

2) The availability of a real-world geospatial datasets providing an empirical basis for pipeline development and evaluation.

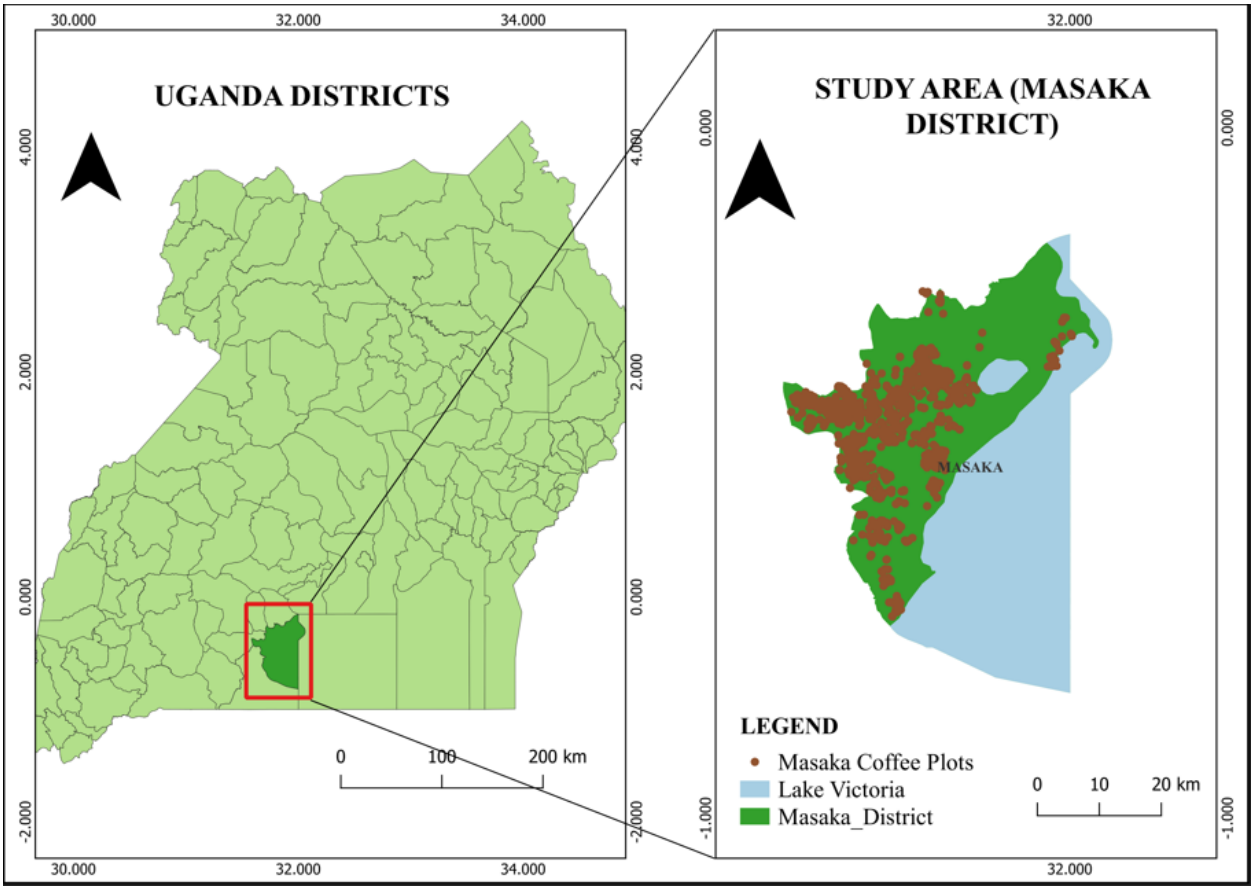


Figure 2: Location of the study area (Masaka District)

3.3 Data Sources and Characteristics

To ensure the integrated pipeline was grounded in real-world conditions, this study utilized a combination of secondary data and reference standards.

3.3.1 Secondary Data

The secondary dataset was acquired from the Uganda Coffee Development Authority (UCDA), consisting of 1,350 individual coffee farm geolocation files from Masaka District, packaged as a single archive (Masaka_Coffee_plots.zip) and uploaded to the Google Colab runtime as the sole manual step in pipeline execution.

3.3.2 Reference Data

In addition to the secondary data, several reference standards and supporting datasets were used during the study. These provided the regulatory and technical framework required for data transformation and compliance alignment.

Table 2: Reference Data

	Data Source	Reference / Standard	Role in the proposed pipeline
1.	Data interchange standard	RFC 7946 (GeoJSON).	Syntactic framework: it was used as the target export syntax to ensure international interoperability with the EUDR-IS
2.	Geodetic standard	EPSG: 4326 (WGS 84)	Spatial reference: it provided the parameters for all coordinate transformations.
3.	Regulatory specifications	EUDR-IS technical specifications	Logical definition: used to dictate mandatory attribute fields and geometric validity rules
4.	Administrative boundary data for Masaka District	UBOS	Area of jurisdiction definition

3.4 Methodology Overview

The study methodology was structured as a two-phase technical framework implemented within a single Google Colab notebook (Geospatial_Integrated_Pipeline. ipynb), executed end-to-end via a single Run All command following the one manual upload step. Phase I (Objective 1) performed a raw dataset audit producing a multi-sheet Excel characterization report and a bridge GeoPackage containing the harmonized master GeoDataFrame, with all EPSG:32636 features already reprojected to WGS84. Phase II (Objective 2) loaded this bridge file and processed it through four sequential compliance modules, producing EUDR submission ready GeoJSON outputs and a full performance analytics report. The two phases were linked through the bridge file, which carried forward both the harmonized geometries and the source-format and CRS provenance metadata required by Objective 2's module logic.

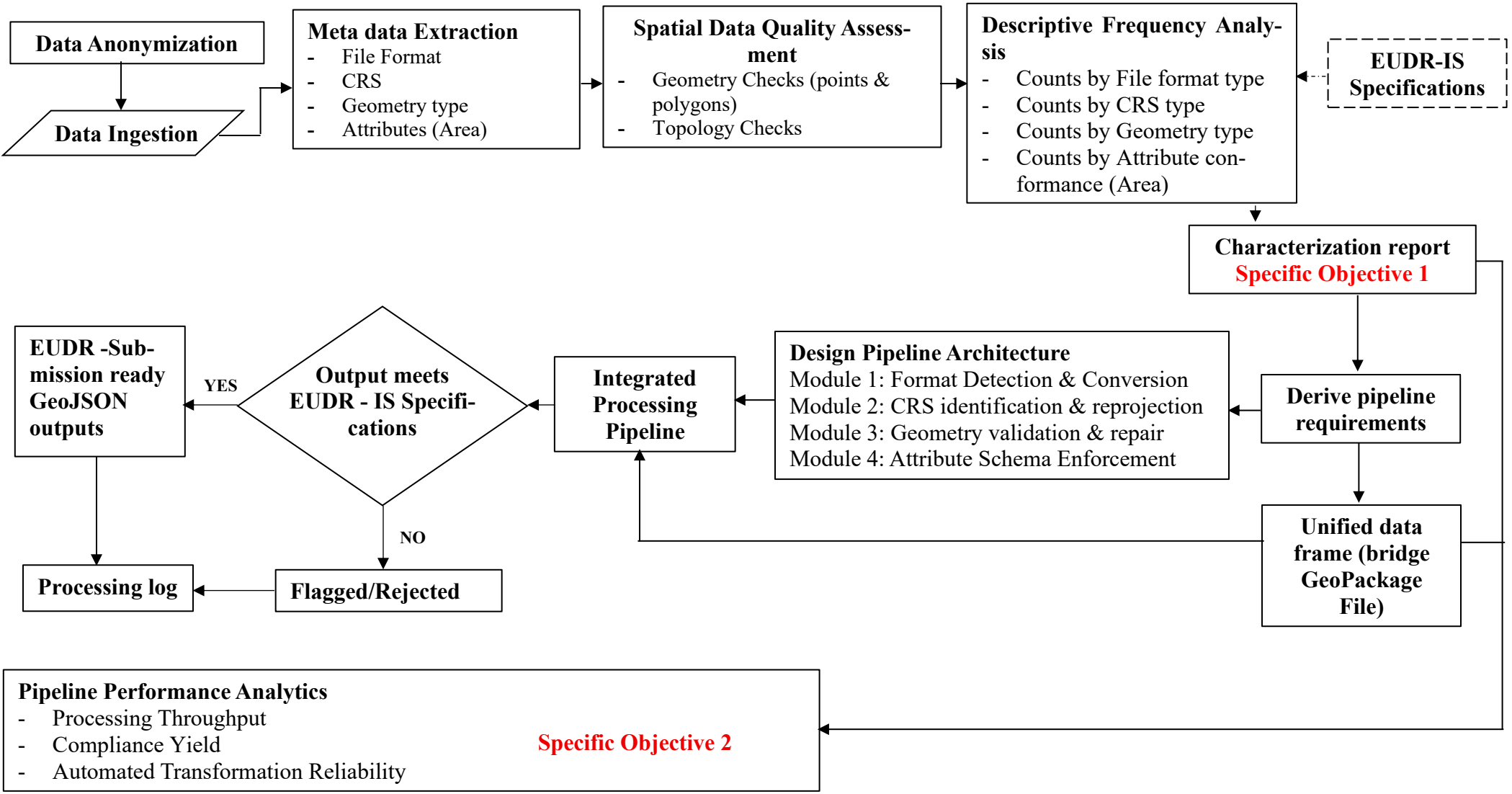


Figure 3:Illustrates the overall methodological workflow adopted in the study.

3.5 Phase I (Specific Objective 1): Data Characterization

A systematic data characterization stage was implemented to establish a technical baseline for the input data, derive functional requirements as well as produce the bridge file consumed by Phase II. The characterization was operationalized through a sequence of notebook cells executing ingestion, file format conversion, geometry type classification, CRS inventory, area analysis and topology auditing.

3.5.1 Data Extraction, Ingestion and Master GeoDataFrame Construction

The raw plot files were pseudonymized to remove personally identifiable information, including farmer names and contact details. The pseudonymized repository was uploaded, extracted and inspected to establish the range of file formats and directory structures present and finally loaded into a single master GeoDataFrame.

Table 3: Back-end logic: Data Extraction, Ingestion and Master GeoDataFrame Construction

Component	Technical Specification
Primary input	Pseudonymized raw file repository (compressed Masaka_Coffee_plots.zip).
Core libraries	`os, glob, zipfile, GeoPandas, Fiona, Pandas, pathlib`
Processing logic	1) Zip extraction and file inventory: The archive was extracted and all file paths were programmatically traversed to create a structured manifest with source_file provenance. 2) Multi-format loading: Native vector formats (GeoJSON, GPKG, Shapefile) were loaded via GeoPandas/Fiona while tabular formats (CSV, XLSX) had WKT geometry columns parsed and converted to spatial features. 3) CRS recording: The original CRS of each source file (crs_original) was recorded prior to any reprojection, preserving provenance for downstream Module 2 logic. 4) EPSG:32636 features were reprojected to WGS84 (EPSG:4326) at this stage using PyProj, so that the master GeoDataFrame carries WGS84 geometry throughout, while crs_original retains the source CRS for audit purposes.
Primary output	Master GeoDataFrame (all features in WGS84) with source_file, source_format, and crs_original provenance fields

3.5.2 Geometry Classification, CRS Inventory, Area Analysis

Each feature was classified by geometry type, the CRS distribution of source files was inventoried and area consistency checks were performed.

Table 4: Back-end logic: Geometry Classification, CRS Inventory and Area analysis

Component	Technical Specification
Primary input	Master GeoDataFrame
Core libraries	`Pandas, GeoPandas, Shapely, os`
Processing logic	1) Geometry type classification: Each feature was assigned a geom_class of Point or Polygon (with Multi-geometries mapped to their respective base class), determining which downstream EUDR rules applied. 2) CRS inventory: The crs_original field recorded during ingestion was tabulated across the dataset to quantify the proportion of files originally referenced in EPSG:4326 versus EPSG:32636. 3) Area consistency: Polygon areas were computed by reprojecting to EPSG:32636 (UTM Zone 36N) and comparing the computed area against the attributed Area_ha field using a tolerance of ±0.001 ha. 4) Large-scale point detection: Point features with attributed Area_ha exceeding 4 ha were flagged as requiring polygon representation under EUDR Article 9.
Primary output	Classified master GeoDataFrame with geom_class, CRS inventory table and area analysis findings table

3.5.3 Spatial Data Quality Assessment

This stage implemented a diagnostic evaluation of polygon geometry validity with topological quality executed as an iterative validation loop against EUDR-IS requirements.

Table 5: Back-end Logic: Spatial data quality assessment

Component	Technical Specification
Primary input	Classified master GeoDataFrame

Component	Technical Specification
Core libraries	`Shapely, Pandas, GeoPandas`
Processing logic	1) Topology validation: Polygon geometries were evaluated using Shapely's is_valid function, aligned with OGC Simple Features standards (ISO 19125), with error types identified via explain_validity() and classified (self-intersection, too few points) 2) Sliver and spike detection: Sliver polygons were identified using a perimeter²/area compactness ratio computed in EPSG:32636, spike artefacts were identified through interior vertex angle analysis.
Decision rule	Invalid geometries were recorded with specific error types. Datasets failing critical quality thresholds (geometry invalidity, topology findings) were flagged for corrective processing in Phase II.
Primary output	Structured findings tables (geometry validity, topology findings, compiled into the multi-sheet Excel characterization report.

3.5.4 Descriptive Frequency Analysis

The final stage of Phase 1 involved quantifying the observed data characteristics together with the detected errors to provide evidence base for Phase II Pipeline design.

Table 6: Descriptive Frequency Analysis

Component	Technical Specification
Primary input	Master GeoDataFrame and Structured findings tables from sections (3.5.1-3.5.3).
Core libraries	`Pandas, GeoPandas (to_file with GeoPackage driver), openpyxl`
Processing logic	1) Excel report compilation: File format distribution, CRS distribution, geometry validity, topology findings and area analysis tables were compiled into a multi-sheet Excel workbook constituting the Objective 1 characterisation report. 2) Bridge file export: The master GeoDataFrame (in WGS84, with reprojected geometries, geom_class, source_format, crs_original, and Area_ha attributes) was exported as a GeoPackage layer (master_plots), forming the bridge file consumed by Phase II.

Component	Technical Specification
	3) Functional requirements derivation: Characterisation findings were translated into a formal functional requirements table, each mapped to a specific module of the Phase II pipeline, and exported as a supporting document.
Primary output	Objective 1 characterisation report (multi-sheet Excel), bridge GeoPackage (master_plots layer), and functional requirements document.

3.6 Phase 2 (Specific Objective 2): Pipeline Design, Implementation and Design

Phase II loaded the Objective 1 bridge file and processed all 1,350 features through four sequential compliance modules followed by an EUDR-IS compliance decision gate and GeoJSON output writer. The pipeline was implemented as a modular, rule-based architecture in which each module addressed a defined class of inconsistencies identified in Phase I. Modules executed sequentially with feature-level conditional branching, with features assigned a REJECTED final_status at any stage were skipped by all subsequent modules, while all transformations were logged non-destructively to a per-feature processing_notes field and a structured processing_log hence ensuring full traceability.

3.6.1 Functional Requirement Definition

Functional requirements for the pipeline were systematically derived from the outputs of the data characterization report (Phase I). Identified inconsistencies were translated into explicit system behaviors that defined the scope and structure of the pipeline.

Table 7: Derived Functional Requirements

Module	Requirement	Phase 1 Basis
Module 1 - Format Detection & Conversion	1) Detect and ingest all five file formats present in the dataset without manual pre-processing.	Five distinct formats identified.
	2) Parse and convert WKT geometry strings from tabular formats (CSV, XLSX) into valid geometry objects.	CSV and XLSX files store geometry as WKT with no native geometry support.
Module 2 - CRS Identification & Reprojection	1) Assign EPSG:4326 as the default CRS to files with no	CSV and XLSX files carry no embedded CRS.

	embedded CRS declaration, validated by coordinate range	
	2) Identify and record the native CRS of every input file prior to any geometric operation	CRS inventory compiled across the dataset (Section 3.5.2)
	3) Reproject all geometries not in EPSG:4326 to EPSG:4326 prior to output generation, in compliance with RFC 7946	GPKG and Shapefile files carry EPSG:32636 (UTM Zone 36N); RFC 7946 mandates WGS84 for GeoJSON (Section 3.5.2)
Module 3 - Geometry Validation & Repair	1) Detect all invalid polygon geometries and attempt best-effort repair using `make_valid()`, logging the repair outcome per feature	Invalid polygon geometries detected and topology integrity issues. (Section 3.5.3)
	2) Detect and log all null or empty geometries, excluding them from compliant output with a documented rejection reason	Null/empty geometry instances detected (Section 3.5.5)
	3) Detect and report all intra-file polygon overlaps	Intra-file polygon overlaps instances detected (Section 3.5.6)
Module 4 - Attribute Schema Enforcement	1) Validate presence of all required EUDR-IS attribute fields (Plot_Id, Farmer_Id, Area_ha, Country, Product) and log missing values per feature	Features found with missing Area_ha (Section 3.5.3)
	2) Recompute Area_ha from polygon geometry (projected to EPSG:32636) and reconcile discrepancies exceeding ±0.001 ha against the attribute value	Polygon features found with area discrepancy exceeding ±0.001 ha (Section 3.5.3)
	3) Flag all point geometry features with Area_ha exceeding 4 hectares as potentially large-scale, appending a compliance flag to the output schema	Point features found with Area_ha > 4 ha (Section 3.5.3)
	4) Produce all compliant outputs as GeoJSON files conforming to RFC 7946, with geometries in EPSG:4326, one feature per Plot_Id, and a	EUDR-IS requires GeoJSON as the required geolocation data format for operator submissions.

	complete EUDR-IS attribute schema

3.6.2 Bridge File Loading, data Pipeline Initialization and work flow

The bridge GeoPackage (master_plots layer) was loaded and its CRS explicitly set to WGS84. Pipeline tracking columns were initialised for every feature: **`compliance_flag`** (defaulted to COMPLIANT), **`processing_notes, output_file, Area_ha_computed`**, and per-module status fields (M1_status through M4_status, final_status), all defaulted to PENDING. Therefore, every feature in the master GeoDataFrame was passed sequentially through four processing modules. Each module evaluated the feature against the criteria established for that module, updated the feature's module-level status and compliance flag accordingly, and appended a corresponding entry to the processing log.

1) Module 1: Format Detection and Conversion

Module 1 performed an initial pass over of a bridge file in GeoPackage (master_plots layer). This included a record of each file with its source format which were five different file formats.

Table 8: Back-end Logic Module 1: Format Detection & Conversion

Component	Technical Specification
Primary input	Bridge GeoPackage (master_plots layer)
Core libraries	**`GeoPandas, Pandas`**
Processing logic	Records in the master bridge file were iteratively processed to identify each file's source format. Format-specific import routines were applied to read the source datasets, after which spatial features and attributes were transformed into a standardized GeoJSON representation. Conversion outcomes and associated metadata were recorded for each file.
Primary output	Master GeoDataFrame with M1_status, compliance flag, and processing notes populated for every feature

2) Module 2: CRS Identification and Reprojection

Module 2 performed a CRS reprojection to EPSG:4326 based on the original CRS recorded for each source file.

Table 9: Back-end Logic Module 2: CRS Identification and Reprojection

Component	Technical Specification
Primary input	Master GeoDataFrame (post-Module 1)
Core libraries	`Pandas`
Processing logic	For each non-rejected feature, the original CRS recorded for its source file (Section 3.5.2) was checked. Features originally and those inferred in EPSG:4326 were confirmed as natively WGS84. Features originally in EPSG:3263 were reprojected to EPSG:4326.
Decision rule	Features that were reprojected and assigned status PASS with no further action required.
Primary output	Master GeoDataFrame with M2_status, compliance flag, and processing notes updated for every non-rejected feature

3) Module 3: Geometry Validation and Repair

Module 3 performed geometry validation, repair and a tiered set of shape-quality corrections, drawing on the validity audit and shape-quality findings from Phase I.

Table 10: Back-end Logic Module 3: Geometry Validation & Repair

Component	Technical Specification
Primary input	Master GeoDataFrame (post-Module 2)
Core libraries	`Shapely, GeoPandas, PyProj`
Processing logic	1. Validity check: each non-rejected feature's geometry was checked for null/empty status and validity using **is_valid.**

Component	Technical Specification
	2. Repair: for invalid geometries other than those classified as "Too Few Points", repair was attempted using **`make_valid()`**. 3. Bounds check: for all non-rejected features, geometry bounds were compared against a reference bounding box for Uganda. 4. Spike removal: for polygon features, vertices with an interior angle below 5° were iteratively removed from each ring, subject to a minimum of three vertices per ring. 5.Sliver repair: for polygon features with a perimeter-squared-to-area ratio exceeding 30.0, a convex hull was computed; if the convex hull's ratio fell within the threshold, the geometry was replaced with its convex hull, otherwise the feature was flagged for review.
Decision rule	Geometries classified as "Too Few Points" were assigned status REJECTED without a repair attempt, as they could not represent a valid plot geometry. Geometries successfully repaired by **`make_valid ()`** were assigned status REPAIRED; those for which repair failed were assigned status REJECTED. Valid geometries were assigned status PASS while those with unresolved geometry and topology issues had their compliance flag escalated to FLAGGED (where not already REJECTED).
Primary output	Master GeoDataFrame with M3_status, repaired geometries, compliance flag, and processing notes updated for every non-rejected feature

4) Module 4: Attribute Schema Enforcement

Module 4 validated the presence of mandatory EUDR-IS attribute fields and reconciled the Area_ha attribute against the geometry-derived area, drawing on the area analysis findings from Phase I.

Table 11: Back-end Logic Module 4: Attribute Schema Enforcement

Component	Technical Specification
Primary input	Master GeoDataFrame (post-Module 3)
Core libraries	`GeoPandas, Pandas`
Processing logic	1. Required field check: each non-rejected feature was checked for the presence of `Plot_Id, Farmer_Id, Area_ha`, Country, and Product. 2. Polygon area reconciliation: for polygon features, the geometry was reprojected to EPSG:32636 and its area computed. Where the Area_ha attribute was missing, it was set to the computed area; where the absolute difference between the attribute and computed area exceeded 0.001 ha, the attribute was updated to the computed value. Polygons with zero computed area were logged as degenerate without triggering an area-mismatch flag. 3. Point area check: for point features, the Area_ha attribute was checked against the 4-hectare large-scale-point threshold.
Decision rule	Features with one or more missing required fields, or point features with `Area_ha` exceeding 4 ha, had their compliance flag set to FLAGGED. Features passing all checks with no missing fields and (for points) within the area threshold were assigned status PASS.
Primary output	Master GeoDataFrame with M4_status, reconciled `Area_ha` and `Area_ha_computed` values, compliance flag, and processing notes updated for every non-rejected feature

3.6.4 EUDR-IS Compliance Decision Gate and Output Generation

Following Module 4, every feature's accumulated compliance_flag and final_status were resolved into one of three terminal classes;

1) `COMPLIANT`, no issues raised across all modules.
2) `FLAGGED`, successfully processed but carrying one or more data quality annotations requiring human review.

3) **`REJECTED`**, unrepairable geometry or unconfirmed CRS, excluded from output.

Thereafter, compliant and flagged features were written as individual GeoJSON files one file per Plot_Id so as to separate output directories for **`eudr_compliant`** and **`eudr_flagged`** respectively.This followed a standardized output schema (Plot_Id, Farmer_Id, Area_ha, Country, Product, compliance_flag, processing_notes, geometry), with Area_ha_computed substituted for Area_ha where available. All output files were written with CRS explicitly set to EPSG:4326 via GeoPandas **`to_file(driver='GeoJSON'),`** conforming to RFC 7946. Rejected features were excluded from all written output but retained in the processing log for audit purposes.

3.6.5 Performance Evaluation

The evaluation emphasized three complementary metrics: Processing throughput, automated transformation reliability and overall compliance yield. Together, these metrics quantified the pipeline's computational efficiency, the reliability of its automated correction routines and its overall effectiveness in producing EUDR submission ready GeoJSON outputs from heterogeneous input data, thereby addressing the study's second research question.

1) Processing Throughput (PT)

Processing Throughput measured the number of features processed per unit of execution time, providing an indication of the computational efficiency and scalability of the pipeline.

It was computed for both the full dataset and for individual pipeline modules, enabling the identification of potential bottlenecks and optimization opportunities.

$$PT = \frac{Total\ Features\ processed}{Combined\ Module\ Execution\ time}$$

2) Automated Transformation Reliability (ATR)

ATR captured the degree to which the pipeline consistently produced valid, correctly transformed outputs. ATR quantified the reliability and consistency of the automation under heterogeneous and potentially noisy input conditions, specifically, for module 2 and module 3.

$$\text{ATR} = \frac{Repaired_{M3} + Reprojected_{M2}}{(Repaired_{M3} + Repair\ Failed_{M3}) + Reprojected_{M2} + Quarantined_{M2}} \times 100$$

3) Overall Compliance Yield (CY)

Overall Compliance Yield assessed the proportion of input features that were retained by the pipeline and written to output as either COMPLIANT or FLAGGED records as opposed to being REJECTED and excluded from all output. It was computed as:

$$CY_{overall} = \frac{(COMPLIANT + FLAGGED)}{Total\ Input\ Features} \times 100$$

And, a stricter variant, Strict Compliance Yield, was also computed as the proportion of features that passed all four modules without requiring any flag or correction:

$$CY_{Strict} = \frac{COMPLIANT}{Total\ Input\ Features} \times 100$$

CHAPTER FOUR: RESULTS AND DISCUSSION

This chapter presents and discusses the findings obtained from the characterization of aggregated coffee farm geolocation datasets from Masaka District and the evaluation of the developed integrated geospatial processing pipeline. The results are organized according to the study objectives. For Objective One, the datasets were assessed against the spatial requirements of the EUDR Information System. The assessment focused on file format conformity, coordinate reference system consistency, geometry validity, topology integrity, attribute consistency and area consistency. The identified inconsistencies were interpreted as technical gaps relative to EUDR geolocation requirements.

For Objective Two, the developed pipeline was evaluated using performance analytics relating to processing efficiency, compliance yield and transformation reliability. Processing efficiency was assessed using module execution times and overall throughput, while compliance yield measured the proportion of datasets successfully transformed into EUDR-geolocation compliant outputs. Transformation reliability was evaluated based on the success of automated correction and standardization operations.

4.1 Overview of the characterization results (Specific objective 1)

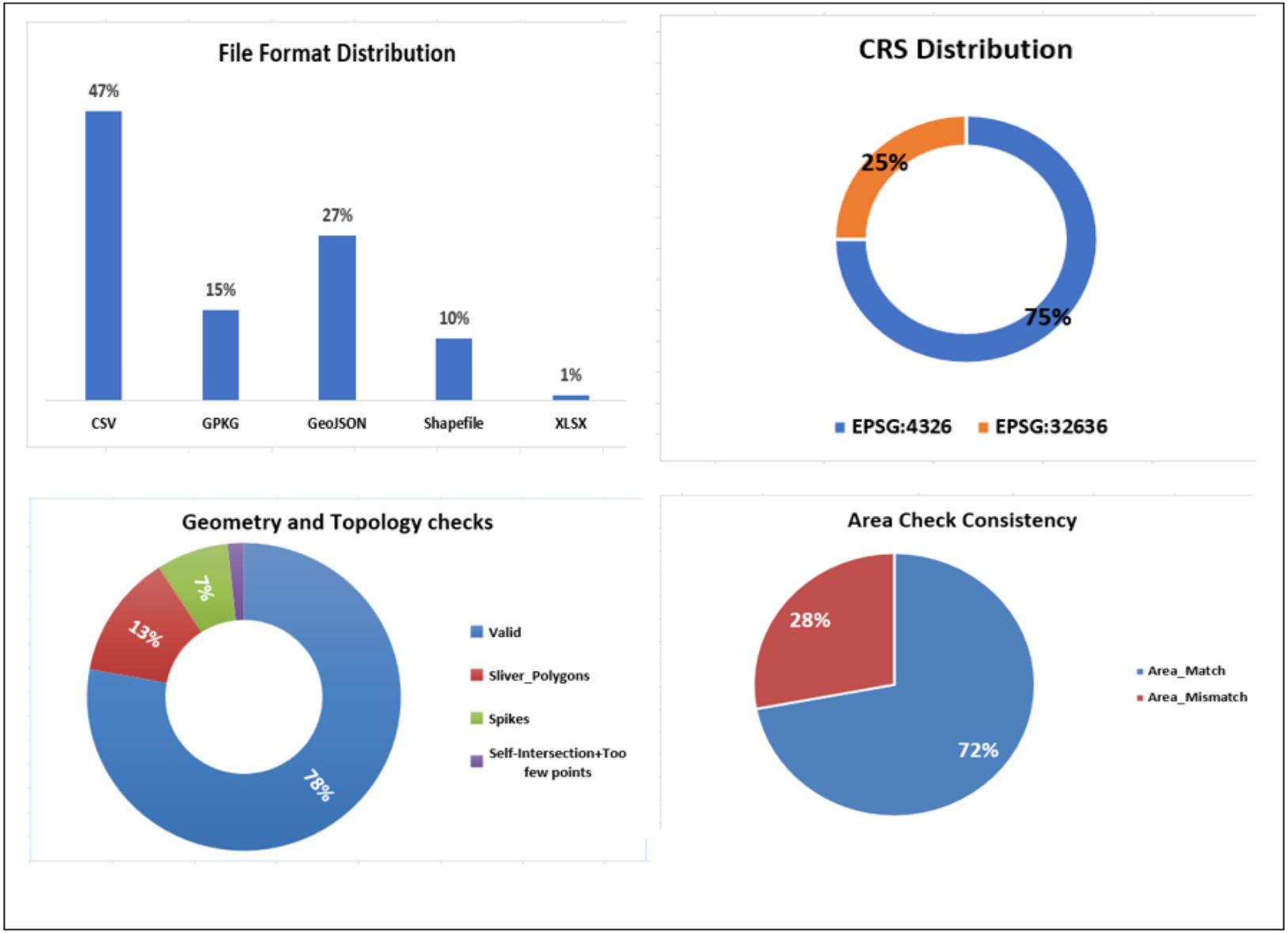


Figure 4: Masaka Coffee Plot Dataset Characterization findings

4.1.1 File format distribution

As shown in Figure 4, five file formats were identified within the aggregated Masaka coffee farm geolocation data, namely CSV (47%), GeoJSON (27%), GeoPackage (15%), Shapefile (10%) and XLSX (1%). The presence of five distinct formats demonstrates considerable heterogeneity in the storage and exchange of geolocation data. Although GeoJSON is the format accepted by the EUDR Information System, only 27% of the datasets were already available in this format. Therefore, approximately 73% of the datasets required format conversion prior to submission. The observed diversity may be attributed to the multi-actor nature of Uganda's coffee supply chain, where exporters, cooperatives and traders are likely to utilize different data collection and management systems. These findings indicate that format standardization represents a major preprocessing requirement for achieving EUDR compliance and ensuring interoperability across the coffee value chain.

4.1.2 CRS distribution

Two coordinate reference systems were identified within the aggregated datasets, namely EPSG:4326 (WGS84) and EPSG:32636 as shown in Figure 4. WGS84 accounted for 75% of the datasets while EPSG:32636 accounted for the remaining 25%. The predominance of WGS84 was partly influenced by the CSV and XLSX datasets, which didn't inherently store CRS metadata. For these datasets, latitude-longitude coordinates recorded in decimal degrees were assumed to be referenced in WGS84 based on their geographic representation and consistency with the study area. From an EUDR compliance perspective, the findings indicate that approximately three-quarters of the datasets already conformed to the coordinate reference system required by the EUDR Information System. Therefore, only 25% of the datasets required CRS transformation prior to submission. This suggests that CRS inconsistency was present but was less pronounced than the file format heterogeneity observed in the dataset.

4.1.3 Geometry validity and topological integrity assessment

The aggregated Masaka coffee farm dataset consisted predominantly of polygon features (93%), while point features accounted for the remaining 7%. Since polygon geometries form the basis for representing larger production units under EUDR requirements, geometry quality assessment focused primarily on the polygon datasets.

Geometry validation revealed that 78% of polygons were valid, while only 22% contained invalid geometries with 2% arising from self-intersections and insufficient vertices. This high validity rate

suggests that the majority of field-mapped boundaries were structurally sound. Nevertheless, the presence of invalid geometries remains important because such features may fail automated compliance checks and require correction before submission.

Among the invalid polygons, 20% exhibited topology integrity issues. Slivers constituted the most common anomaly (13%), followed by spikes (7%). These anomalies are commonly associated with digitization, boundary editing and geospatial preprocessing operations. Although they do not necessarily render geometries invalid, they can affect area calculations and spatial accuracy. Consequently, topology correction remains an important step in ensuring reliable EUDR-compliant geolocation data.

4.1.4 Attribute schema-Area consistency check

As part of the attribute schema assessment, declared plot areas were compared against areas computed from polygon geometries. A tolerance threshold of ±0.0001 hectares was applied to account for minor numerical differences arising from coordinate precision, geometric representation and computational rounding. Using this criterion, 902 polygons were classified as area matches while 350 polygons exhibited area discrepancies exceeding the defined tolerance. This implies that approximately 28% of the assessed polygons contained area inconsistencies.

One possible explanation for these discrepancies is that reported area values may not always correspond to the mapped coffee plot extent. For example, farmers may report the size of the entire farm area while only a portion of the land is represented as a coffee production plot. Such inconsistencies may also arise from manual data entry errors or changes made to plot boundaries after area information had been recorded.

From an EUDR compliance perspective, area consistency is particularly important because plot size influences the required representation of production units. Supporting this observation, eight features were identified with declared areas exceeding four hectares while being represented as point geometries. Since EUDR specifications require production units larger than four hectares to be represented as polygons, these records constitute direct compliance violations. The findings therefore highlight the importance of attribute validation alongside geometric validation in EUDR geolocation data preparation.

4.2 Design, Implementation and Performance Evaluation of the Integrated Geospatial Processing Pipeline (Specific Objective 2)

4.2.1 Pipeline Architecture Overview

The pipeline was architectured as a four-module sequential processing chain, with each module addressing a distinct class of data quality non-conformance identified in Objective 1. The modules and their primary functions are summarized in Table 12.

Table 12: Pipeline Module Architecture and Primary Functions

Module	Name	Primary Function
M1	Format Detection and Conversion	Ingest all five source formats (CSV, XLSX, GeoJSON, GPKG, Shapefile); parse WKT geometry from tabular formats; build unified master GeoDataFrame
M2	CRS Identification and Reprojection	Identify native CRS of every file; reproject EPSG:32636 features to WGS84 (EPSG:4326); quarantine features with unknown CRS
M3	Geometry Validation and Repair	Validate OGC geometry rules; apply Shapely make_valid() for repairable errors; reject unrepairable features (too few points); flag out-of-bounds coordinates
M4	Attribute Schema Enforcement	Validate required EUDR-IS fields; reconcile area mismatches using UTM-projected computation; flag large-scale points exceeding 4 ha threshold

4.2.2 Processing Throughput

The results reveal that all four modules were executed within a relatively short time frame, with execution times ranging from approximately 0.3 to 10 seconds. Attribute schema validation recorded the highest processing time (8.896 seconds), followed by CRS reprojection (5.006 seconds), and geometry repair (2.964 seconds) and format conversion (0.333 seconds). This can be attributed to the fact that attribute validation involved multiple compliance checks including area consistency verification and EUDR-specific 4-hectare rule evaluation, resulting in a greater computational workload than the other modules. Generally, the relatively small differences in execution time across the four modules indicate a balanced workload distribution within the integrated pipeline. No single module emerged as a major computational bottleneck, suggesting that the workflow architecture effectively distributed processing tasks across the standardisation, validation and compliance assessment stages.

Additionally, it should be noted that the execution time of individual modules and the overall processing throughput 78.49 features per second were partly dependent on network conditions. Because the workflow was implemented in a cloud-based execution environment, variations in internet bandwidth and latency influenced the computational performance metrics reported in this study.

4.2.3 Automated Transformation Reliability (ATR)

Transformation reliability was evaluated by comparing attempted corrective operations with their corresponding outcomes. The reprojection and geometry repair operations were executed, achieving a 87% success rate. Features identified as having insufficient vertices for valid polygon construction were not subjected to automated correction and were instead rejected in accordance with the quality control rules defined in the processing framework.

These results demonstrate that the pipeline reliably executed all supported corrective procedures while consistently enforcing exclusion criteria for non-recoverable geometries. The findings indicate that automated transformation and validation mechanisms can substantially reduce manual intervention during EUDR compliance preparation, while maintaining the integrity of the resulting geospatial dataset.

4.2.4 Compliance Yield

The pipeline produced compliant outputs (submission ready GeoJSON) 1202 of the 1350 assessed features, corresponding to a compliance yield of approximately 89%. A further 147 datasets were flagged for review while only a single dataset was rejected. The high proportion of compliant outputs demonstrates the effectiveness of the integrated workflow in transforming heterogeneous geolocation datasets into EUDR submission ready products. Furthermore, the extremely low rejection rate indicates that most identified inconsistencies were either corrected automatically or successfully isolated through the flagging mechanism.

CHAPTER FIVE: CONCLUSION AND RECOMMENDATIONS

5.1 Conclusion

5.1.1 Conclusion for Specific Objective 1: Data Characterization

The study established that the aggregated coffee farm geolocation datasets exhibited substantial technical heterogeneity that limited their direct suitability for EUDR-IS submission. The datasets were distributed across multiple file formats, including CSV, GeoJSON, GeoPackage, Shapefile and XLSX, while two coordinate reference systems (EPSG:4326 and EPSG:32636) were identified. Geometry validation further revealed the presence of invalid polygons resulting from self-intersections and insufficient vertices, while topological assessment identified slivers and spikes that could compromise spatial integrity. Additional inconsistencies were observed in attribute information through area mismatches and the presence of point features exceeding the EUDR four-hectare threshold. Variations in coordinate precision were also detected. Collectively, these findings demonstrated that although portions of the datasets were submission already, significant technical gaps existed relative to EUDR-IS requirements, thereby necessitating systematic preprocessing and standardization before submission.

5.1.2 Conclusion for Specific Objective 2: Pipeline Development and Evaluation

The implementation and evaluation of the integrated geospatial pipeline developed within the Google Colab environment successfully addressed the baseline inconsistencies identified in the first objective. Processing data at a throughput of 78.49 features per second, the pipeline demonstrated high efficiency, effectively isolating computational bottlenecks within the initial file format conversion stage while executing coordinate transformations, geometric repair and schema audits. With a complete transformation success rate across all attempted CRS reprojections and geometric repairs, the pipeline proved to be highly reliable. Most importantly, the pipeline converted a flawed district dataset into an 89.04% submission-ready GeoJSON assets, while safely flagging data anomalies for review and filtering out unrecoverable file corruption. This establishes automated programmatic pipelines as a viable and scalable replacement for manual desktop GIS workflows in smallholder supply chains.

5.2 Recommendations

1) Implement Upstream Field Validation

The characterization of the aggregated coffee farm geolocation datasets revealed inconsistencies in file formats, coordinate reference systems, geometry validity, area attributes and coordinate precision. To reduce the occurrence of these issues, organizations involved in geolocation data collection should implement upstream field validation procedures. Validation rules relating to co-ordinate precision, CRS standardization, plot size requirements and geometry quality should be enforced during data capture to minimize the propagation of errors into downstream compliance workflows.

2) Adopt Automated Geospatial Compliance Processing Workflows

The performance evaluation demonstrated that the integrated pipeline successfully transformed heterogeneous geolocation datasets into EUDR-submission ready GeoJSON outputs with high processing efficiency and transformation reliability. Coffee exporters and compliance managers should therefore adopt automated geospatial processing workflows as part of routine compliance operations to improve consistency, reduce manual effort and support large-scale preparation of EUDR submission datasets.

3) Develop a QGIS-Based Implementation of the Pipeline

Although the developed solution was successfully implemented in a Google Colab environment, future implementation should focus on developing a QGIS plugin that encapsulates the pipeline functionality within a user-friendly graphical interface. Such an implementation would improve accessibility among GIS practitioners and compliance personnel while facilitating wider operational adoption across the coffee sector.

4) Future Research

Future research should extend the current pipeline beyond data standardization and validation by incorporating automated deforestation-risk assessment capabilities using Earth Observation data. Additionally, direct interoperability with the EUDR Information System should be explored to support end-to-end compliance workflows from data preparation to submission.

REFERENCES

Afriyie-Kraft, L., Zabel, A., & Breu, T. (2026). Dealing with the cost of EUDR compliance data and potential monetisation: Opportunities and risks. *Forest Policy and Economics*, *184*, 103709. https://doi.org/10.1016/j.forpol.2026.103709

Alam, K., Roy, B., & Serebrenik, A. (2023). *Reusability Challenges of Scientific Workflows: A Case Study for Galaxy*.

Al-Yadumi, S., Xion, T. E., Wei, S. G. W., & Boursier, P. (2021). Review on Integrating Geospatial Big Datasets and Open Research Issues. *IEEE Access*, *9*, 10604–10620. https://doi.org/10.1109/ACCESS.2021.3051084

Berihun, T. (2026). *EU deforestation regulation and Ethiopian smallholder coffee farmers: Awareness, compliance capacity, and potential impacts*.

Biljecki, F., Chow, Y. S., & Lee, K. (2023). Quality of crowdsourced geospatial building information: A global assessment of OpenStreetMap attributes. *Building and Environment*, *237*. https://doi.org/10.1016/j.buildenv.2023.110295

Brazeal, R., Eng, P., & Surv, P. (2016). *Issues with Extending a UTM Zone*.

Clement, A., & Joel, B. (2022). *Review and Analysis of policies, strategies, attitudes, practices and institutions for strengthening coffee agroforestry in Uganda*.

Du, J. (2026). Opportunities and Challenges of IS in the Era of Artificial Intelligence-Taking precision agriculture as an example. In *International Journal of World Economic Research* (Vol. 1, Number 1).

European Commission. (2023, May 31). *Regulation (EU) 2023/1115 of the European Parliament and of the Council on deforestation-free products*. Official Journal of the European Union. https://eur-lex.europa.eu/eli/reg/2023/1115/2025-12-26

European Council. (2025, December 4). *EU deforestation law: Council and Parliament reach a deal on targeted revision*. https://www.consilium.europa.eu/en/press/press-releases/2025/12/04/eu-deforestation-law-council-and-parliament-reach-a-deal-on-targeted-revision/

Fortin, J., Meacham, S., Rissolo, D., Rosland, A., & Dorshow, W. (2025). Bridging Photogrammetry and GIS Through Workflow Automation: A Scalable Approach for Spatially-Linked 3D Data in Underwater Cultural Heritage. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*, *48*(M-9–2025), 497–502. https://doi.org/10.5194/isprs-archives-XLVIII-M-9-2025-497-2025

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). DESIGN SCIENCE IN INFORMATION SYSTEMS RESEARCH 1. In *Design Science in IS Research MIS Quarterly* (Vol. 28, Number 1).

Hu, T., Liu, T., Divyacharan Jarugumalli, V. S., Cheng, S., & Deng, C. (2025). FAIR principles in workflows: A GIScience workflow management system for reproducible and replicable studies. *International Journal of Applied Earth Observation and Geoinformation*, *138*. https://doi.org/10.1016/j.jag.2025.104477

Internation Labour Organization. (n.d.). *Mapping the coffee value chain in Uganda Chains project*.

Khan, F. Z., Soiland-Reyes, S., Sinnott, R. O., Lonie, A., Goble, C., & Crusoe, M. R. (2019). Sharing interoperable workflow provenance: A review of best practices and their practical application in CWLProv. *GigaScience*, *8*(11). https://doi.org/10.1093/gigascience/giz095

Kleinschmit, D., Blum, M., Brockhaus, M., Karambiri, M., Kröger, M., Ramcilovic-Suominen, S., & Reinecke, S. (2023). *Key Logics of International Forest Governance and SDG 15*.

Kristin, D., Naoko, T., Thomas, W., Caroline, M., Serena, F., & Astrid, A. (2022). Halting deforestation from agricultural value chains: the role of governments. In *Halting deforestation from agricultural value chains: the role of governments*. FAO. https://doi.org/10.4060/cc2262en

Lambin, E. F., & Furumo, P. R. (2026). *Annual Review of Environment and Resources Deforestation-Free Commodity Supply Chains: Myth or Reality?* https://doi.org/10.1146/annurev-environ-112321

Mansourian, A., & Oucheikh, R. (2026). Bridging natural language and GIS: a multi-agent framework for LLM-driven autonomous geospatial analysis. *International Journal of Digital Earth*, *19*(1). https://doi.org/10.1080/17538947.2026.2633849

Masson, M., Cayèré, C., Bessagnet, M. N., Sallaberry, C., Roose, P., & Faucher, C. (2022). An ETL-like platform for the processing of mobility data. *Proceedings of the ACM Symposium on Applied Computing*, 547–555. https://doi.org/10.1145/3477314.3507057

Musinguzi, M., Stevens, T.-T. S., & Bax, G. (2012). GIS Data Interoperability in Uganda ∗. *International Journal of Spatial Data Infrastructures Research*, *7*, 488–507. https://doi.org/10.2902/1725-0463.2012.07.art22

Othman, N., & Idris, H. N. (2023). *CHALLENGES AND OPPORTUNITIES OF GEOLOCATION APPLICATION IN AGRICULTURE: A SCOPING REVIEW* (Number 17).

Radosevic, N. (2023). *Geospatially Enabled Scientific Workflows for Reproducible Environmental Modeling and Decision Support*.

Schlögl, M., Waltersdorfer, L., Regner, P., Siposova, A., & Brenning, A. (2026). Overcoming barriers to reproducibility in geoscientific data analysis: Challenges and practical implementation strategies. *Environmental Modelling & Software*, *200*, 106962. https://doi.org/10.1016/j.envsoft.2026.106962

Simonetti, D., Degreve, L., Bourgoin, C., Achard, F., & Colditz, R. (2025). *A guide to visualize and interact with GFC2020 using the JRC IMPACT Toolbox and other applications*

Geospatial Analysis Tools supporting the Risk Assessment of the Regulation on deforestation-free supply chains. Joint Research Centre (JRC), the European Commission's science and knowledge service. https://doi.org/10.2760/5075197

Stewart, A. J., Robinson, C., & Banerjee, A. (2025). *Geospatial Machine Learning Libraries*. http://arxiv.org/abs/2510.02572

UCDA. (2020). *COFFEE SUB-SECTOR STRATEGY (2020/21 – 2024/25).*

UCDA. (2025). *MONTHLYCOFFEEREPORT-JULY2025*.

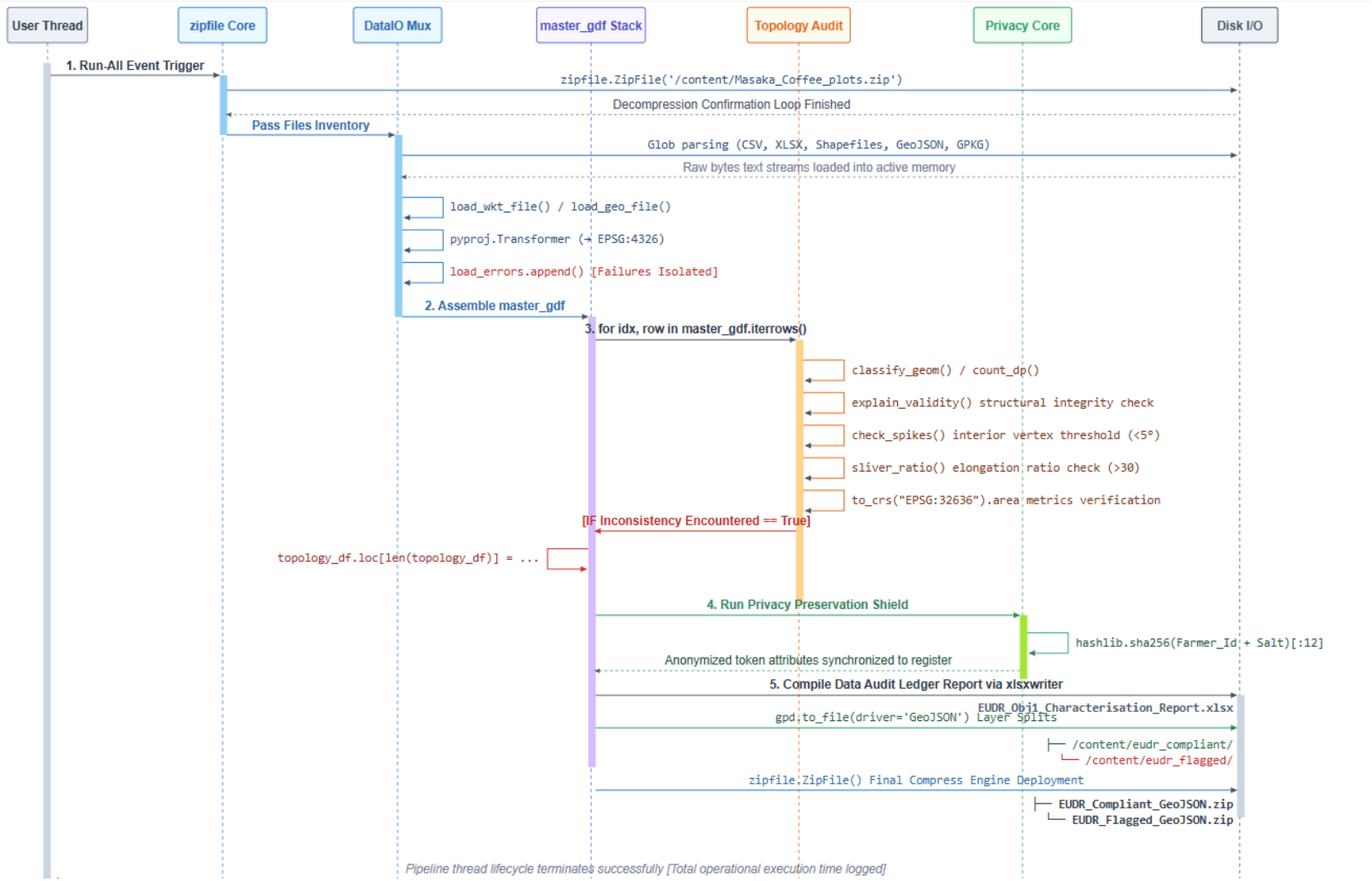
Verma, R. (2024). *Spatial Data Capture in GIS: A Review*. www.allmultidisciplinaryjournal.com

Wardell, D. A., Piketty, M.-G., Lescuyer, G., & Pacheco, P. (2021). *February 2021 Reviewing initiatives to promote sustainable supply chains The case of forest-risk commodities*.

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J. W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). Comment: The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*. https://doi.org/10.1038/sdata.2016.18

APPENDICES

Appendix A: Runtime Execution Sequence Diagram



Appendix B: Sample Audit Report

Plot_Id	Source_Format	Source_File	Record_Type	Validity_Status	Error_Type	Validity_Reason
F-0001	CSV	F-0001.csv	VALIDITY	Valid	Valid	Valid Geometry
F-0004	CSV	F-0004.csv	VALIDITY	Valid	Valid	Valid Geometry
F-0005	CSV	F-0005.csv	VALIDITY	Valid	Valid	Valid Geometry
F-0007	CSV	F-0007.csv	VALIDITY	Invalid	Too Few Points	Too few points in geometry component[31.740145 -0.656977]
F-0009	CSV	F-0009.csv	VALIDITY	Valid	Valid	Valid Geometry
F-0011	CSV	F-0011.csv	VALIDITY	Valid	Valid	Valid Geometry
F-0013	CSV	F-0013.csv	VALIDITY	Valid	Valid	Valid Geometry
F-0015	CSV	F-0015.csv	VALIDITY	Valid	Valid	Valid Geometry

Appendix C: Sample CRS Inventory Result

Plot_Id	Source_Format	Source_File	CRS	CRS_Note
F-0954	CSV	F-0954.csv	EPSG:4326	Inferred — no embedded CRS; decimal-degree coordinates confirm WGS84
F-1121	CSV	F-1121.csv	EPSG:4326	Inferred — no embedded CRS; decimal-degree coordinates confirm WGS84
F-0946	XLSX	F-0946.xlsx	EPSG:4326	Inferred — no embedded CRS; decimal-degree coordinates confirm WGS84
F-0947	GeoJSON	F-0947.geojson	EPSG:4326	Embedded as urn:ogc:def:crs:OGC:1.3:CRS84 (≡ EPSG:4326)
F-1161	GeoJSON	F-1161.geojson	EPSG:4326	Embedded as urn:ogc:def:crs:OGC:1.3:CRS84 (≡ EPSG:4326)
F-1186	GPKG	F-1186.gpkg	EPSG:32636	Embedded in file
F-1105	Shapefile	F-1105.shp	EPSG:32636	Embedded in file
F-1141	Shapefile	F-1141.shp	EPSG:32636	Embedded in file

Appendix D: Sample Area Consistency check

Plot_Id	Source_Format	Geometry_Class	Area_ha_Attribute	Area_ha_Compute	Area_Difference_ha	Area_Flag	Area_Flag_Detail
F-0723	GeoJSON	Polygon	0.17	0.17	0	OK	
F-0729	GeoJSON	Polygon	0.21	0.21	0	OK	
F-0730	GeoJSON	Polygon	0.44	0.44	0	OK	
F-0733	GeoJSON	Polygon	0.32	0.32	0	OK	
F-0734	GeoJSON	Polygon	0.23	0.23	0	OK	
F-0735	GeoJSON	Polygon	0.22	0.22	0	OK	
F-0737	GeoJSON	Polygon	1.14	1.14	0	OK	
F-0738	GeoJSON	Polygon	0.34	0.33	0.01	AREA_MISMATCH	Attr=0.34 ha | Computed=0.330000 ha | Diff=0.010000 ha
F-0740	GeoJSON	Polygon	0.22	0.22	0	OK	
F-0742	GeoJSON	Polygon	1.36	1.36	0	OK	
F-0744	GeoJSON	Polygon	0.61	0.61	0	OK	
F-0745	GeoJSON	Polygon	1	0.92	0.08	AREA_MISMATCH	Attr=1.0 ha | Computed=0.920000 ha | Diff=0.080000 ha

Appendix E; Process Throughput

Metric	Value	Unit	Description
Total Input Features	1350	features	All features entering pipeline
Processing Time	17.199	seconds	Sum of all 4 module times
Overall Throughput	78.49	feat/sec	Total features / processing time
Module 1 - Format Detection	0.333	seconds	Module wall-clock time
Module 2 - CRS Reprojection	5.006	seconds	Module wall-clock time
Module 3 - Geometry Validation	2.964	seconds	Module wall-clock time
Module 4 - Attribute Enforcement	8.896	seconds	Module wall-clock time
GeoJSON Write	10.982	seconds	Module wall-clock time

Appendix F: Processing log

Source_Forma	Source_File	Is_Duplicat	M1_Status	M2_Status	M3_Status	M4_Status	Final_Status	Compliance_Fla	Processing_Notes
CSV	F-0007.csv	FALSE	PASS	PASS	REJECTED	SKIPPED	REJECTED	REJECTED	M1: format=CSV geom=MultiPolygon - PASS | M2: CRS=EPSG:4326 natively - PASS | M3: Too Few Points - geometry is fundamentally unrepresentable and cannot satisfy EUDR-IS Art.3(d) - rejected without repair attempt
GeoJSON	F-0046.geojson	FALSE	PASS	PASS	PASS	PASS	COMPLIANT	COMPLIANT	M1: format=GeoJSON geom=Polygon - PASS | M2: CRS=EPSG:4326 natively - PASS | M3: Geometry valid - PASS | M4: Area mismatch attr=1.2800 ha computed=1.2761 ha - updated to computed | M4: Schema valid - PASS
GeoJSON	F-0298.geojson	FALSE	PASS	PASS	PASS	PASS	COMPLIANT	COMPLIANT	M1: format=GeoJSON geom=Polygon - PASS | M2: CRS=EPSG:4326 natively - PASS | M3: Geometry valid - PASS | M4: Area mismatch attr=1.2400 ha computed=1.2388 ha - updated to computed | M4: Schema valid - PASS
GPKG	F-0002.gpkg	FALSE	PASS	REPROJECTED_IN_OBJ1	PASS	PASS	COMPLIANT	COMPLIANT	M1: format=GPKG geom=Polygon - PASS | M2: Source was EPSG:32636 - reprojected to WGS84 in Objective 1 - geometry confirmed WGS84 - PASS | M3: Geometry valid - PASS | M4: Area mismatch attr=0.1400 ha computed=0.1361 ha - updated to computed | M4: Schema valid - PASS
GPKG	F-1224.gpkg	FALSE	PASS	REPROJECTED_IN_OBJ1	PASS	PASS	COMPLIANT	COMPLIANT	M1: format=GPKG geom=Polygon - PASS | M2: Source was EPSG:32636 - reprojected to WGS84 in Objective 1 - geometry confirmed WGS84 - PASS | M3: Geometry valid - PASS | M4: Area mismatch attr=0.2600 ha computed=0.2571 ha - updated to computed | M4: Schema valid - PASS
Shapefile	F-0092.shp	FALSE	PASS	REPROJECTED_IN_OBJ1	PASS	PASS	COMPLIANT	COMPLIANT	M1: format=Shapefile geom=Polygon - PASS | M2: Source was EPSG:32636 - reprojected to WGS84 in Objective 1 - geometry confirmed WGS84 - PASS | M3: Geometry valid - PASS | M4: Area mismatch attr=0.4300 ha computed=0.4238 ha - updated to computed | M4: Schema valid - PASS
Shapefile	F-0296.shp	FALSE	PASS	REPROJECTED_IN_OBJ1	PASS	PASS	COMPLIANT	COMPLIANT	M1: format=Shapefile geom=Polygon - PASS | M2: Source was EPSG:32636 - reprojected to WGS84 in Objective 1 - geometry confirmed WGS84 - PASS | M3: Geometry valid - PASS | M4: Area_ha verified (0.3204 ha) | M4: Schema valid - PASS