



Università
Ca'Foscari
Venezia

**Master Degree in Data Analytics for Business and
Society**

**Supplier Risk Assessment Under
Uncertainty:
Developing a Structured Evaluation
Approach**

Student: Asma Taba

Student ID: 98754

Supervisor: Prof.Francesco Rullani

Co-supervisor: Mr.Francesco maria Rizzi

Academic Year 2025/2026

Abstract

The analysis of supplier risk has become more complicated due to increased complexity of the supply chain, increased level of disruption and uncertainty. Currently available frameworks mostly operate with data on quality, costs and delivery. They do not produce a complete picture of the risks involved. In addition to that, one of the problems faced by companies in assessing their risks is obtaining the relevant information and incorporating it into the assessment. This thesis explores an alternative methodology for conducting data-driven supplier risk assessment in the described situation. This includes analysis of combining internal quality data, logistics data and selected external data sources and determining whether logistics and external data bring value beyond the internal quality data.

In order to do that, the thesis uses a case of Altromercato organization. The data used include various data sources related to operations, quality and logistics metrics extracted from the company's shipment records as well as country-level risk indicators from the external data sources. The work follows the CRISP-DM framework and involves stages of data preparation, modeling and evaluation.

The framework evaluates the value of a particular data source through the modeling scenarios. The statistical screening is used to determine the predictors associated with shipment results. The regression models and Isolation Forest algorithm are applied to explore potential patterns that indicate risk cases.

The main finding is that the value of a data source does not depend only on its relevance, but also on its availability and consistency. Instead of considering data constraints as peripheral aspects, this thesis focuses on them as core elements of a supplier risk framework. The key contribution of this work is the proposed methodology for organizing and combining various data sources of the supplier risk in the real world.

Contents

1	Literature Review	9
1.1	Supplier Risk: Foundations and Evolution	9
1.1.1	Definition and Scope	9
1.1.2	Historical Development of Supplier Risk Management	9
1.1.3	Categories of Supplier Risk	10
1.1.4	Supplier Risk Assessment and Supplier Selection	12
1.2	Uncertainty in Supply Chains	12
1.2.1	Sources of Uncertainty	12
1.2.2	Supply Chain Vulnerability Under Uncertainty	13
1.3	Supplier Risk Assessment Approaches	13
1.3.1	Traditional Approaches	13
1.3.2	Quantitative Models	14
1.3.3	Data-Driven and Anomaly Detection Approaches	14
1.3.4	Limitations of Existing Methods	15
1.3.5	Toward a Universal Supplier Risk Assessment Framework	15
1.4	Risk Indicators and Evaluation Challenges	17
1.5	Data as a Recurring Challenge	18
1.6	Research Gap	19
1.7	Conclusion	19
2	Research Methodology and Framework Design	21
2.1	Objectives and Research Questions	21
2.1.1	Research main Objectives	21
2.1.2	Research Question	21
2.2	Research Methodology	21
2.3	Business Understanding and Proposed Approach	23
2.4	Data Understanding and Preparation	24
2.4.1	Internal Data	24
2.4.2	Logistics Data	25
2.4.3	External Data	26

2.4.4	Data Integration	26
2.4.5	Data Cleaning and Preparation	27
2.5	Framework Design	27
2.5.1	Evaluation Strategy	28
2.6	Conclusion	29
3	Data Extraction and Preparation	30
3.1	Internal Data Extraction	30
3.1.1	Source of Quality 1 Data Extraction	31
3.1.2	Additional Internal Information	32
3.2	Logistics Data Formulation	34
3.3	External Data Extraction	35
3.4	Data Preparation	37
3.4.1	Structure of the Consolidated Dataset	37
3.4.2	Data Quality and Cross-Source Reconciliation	37
3.4.3	Comparability of Additional Internal Data	38
3.4.4	Individual Significance Testing	38
3.4.5	Correlation Screening and Redundancy Reduction	40
3.4.6	Multicollinearity Check	42
3.4.7	Feature Engineering and Modeling Dataset	42
3.5	Conclusion	42
4	Modeling and Results	44
4.1	Introduction	44
4.2	Initial Exploratory Composite Risk Scoring	44
4.2.1	Entropy Weight Method	45
4.2.2	Evaluation Against Observed Outcome	45
4.3	Unsupervised Modeling: Isolation Forest	46
4.3.1	Initial Exploration	46
4.3.2	Scenario-Based Analysis	46
4.3.3	Final Refinement	48
4.3.4	Leakage-Safe Sensitivity Analysis	48
4.4	Supervised Modeling: Logistic Regression	49
4.4.1	Model Specification and Missing-Data Treatment	50
4.4.2	Regression Results	50
4.4.3	Holdout Evaluation	51
4.5	Comparison of Modeling Approaches	53
4.6	Conclusion	55

5	Discussion	57
5.1	Theoretical Implications	57
5.1.1	Answering the research question	57
5.1.2	Interpretation of the Regression Coefficients	58
5.1.3	Comparison with Literature	59
5.2	Practical and Managerial Implications	60
5.3	Limitations	62
6	Conclusion and Future Research	64
6.1	Conclusion	64
6.2	Theoretical and Methodological Contribution	65
6.3	Practical and Analytical Reflections	67
6.3.1	Practical Implications for Altromercato	67
6.3.2	Rationale Behind the Progression of Analytical Approaches	68
6.4	Limitations	69
6.5	Future Research	70
	Annex	74
A	Supplementary Data Extraction and Analysis Material	75
B	Variable Definition and List of Abbreviations	80

List of Figures

2.1	CRISP-DM Steps (extracted from https://medium.com/).	22
3.1	Internal quality reports across a single origin (Peru)	33
3.2	Fairtrade Risk Map. Source: https://riskmap.fairtrade.net/	36
3.3	Significance test of candidate variables against shipment status	39
3.4	Correlation heatmap of the final regression feature set	41
4.1	Isolation Forest ROC-AUC across scenarios	47
4.2	sensitivity check by scenario 70/30 split	49
4.3	Holdout ROC curve, median-imputed logistic regression	52
4.4	Confusion matrix	52
5.1	Current versus desired state of each information axis Generated by Claude (Anthropic)	61

List of Tables

1.1	Main Categories of Supplier Risk	11
1.2	External and Internal Risk Indicators in Supply Risk Assessment	17
2.1	Scenarios used to evaluate each dimension	28
3.1	Additional internal sources.	33
3.2	Logistics indicators	35
3.3	Fairtrade Risk Map ratings.	36
3.4	Main datasets list.	37
3.5	Screening results against shipment status.	40
4.1	Results for the six-scenario Isolation Forest.	47
4.2	Logistic regression coefficients (N=113, 28 Rejected)	51

4.3	Holdout classification metrics at the default 0.5 probability threshold. . .	53
4.4	Comparison of the three modeling approaches.	54
5.1	Logistic regression coefficients.	58
A.1	Main information extracted from Source of Quality 1 reports.	76
A.2	Not significant Variables	77
A.3	Variance inflation factors	78
A.4	Isolation Forest results	78
A.5	Feature-selection frequency across repeated cross-validation	79
B.1	Principal variables used in the final modeling dataset.	80
2	List of Abbreviations	81

List of Equations

Logistic regression model	49
Odds ratio	50

General Introduction

Supply chains today experience greater complexity, higher levels of uncertainty, and more disruptions than before. Earlier studies considered suppliers' risks by using various parameters such as internal processes, logistics performance measures, environmental factors, and decision-support approaches based on the data analysis. However, there is an identified gap in the literature: information related to supplier risks is disparate, incomplete, and fragmented among various sources making any comparison or integration not feasible.

The reason for carrying out this study is the existence of this gap. Based on the review of literature concerning supplier risks, supply chain uncertainty, data-driven approaches to risk assessment, and limitations associated with data availability and validity, the current work attempts to develop and empirically examine a methodology for combining multiple information sources in the supplier risk assessment process. Specifically, this paper focuses on the analysis of the role that internal quality data, logistics information, and country-level variables may play in risk assessment process and what contribution each type of information makes in identifying risky shipments.

The idea of the research is based on the fact that a single information source or a single technique for analyzing it is insufficient to assess supplier risks. The concept of supplier risk assessment is considered a multidimensional task where the value of information is determined not only by its importance theoretically but also by its quality and availability. Therefore, data preparation and data integration techniques are combined with different analytical tools in order to examine the performance of various information sources in the context of supplier risks.

The empirical case study: This research approach is examined through a real-world case study, using data provided by Altromercato [16], a fair-trade organization that has operated in Italy since 1988 and was among those involved in building the country’s fair-trade movement. Altromercato imports and sells a range of fair-trade goods through sourcing partnerships built on continuity and long-term cooperation. This research focuses only on Cocoa product, as these records provided the information needed to apply that research. The results therefore represent a cocoa-specific application and not an analysis of Altromercato’s entire products.

Data confidentiality: For confidentiality reasons, this thesis does not disclose the identity of the suppliers, nor the external quality-control providers. These companies are anonymized by using generic identifiers: Supplier 1, Supplier 2, while the quality-control providers appear as Source of Quality 1, Source of Quality 2, and Source of Quality 3. Country-level information is retained since it matters to the analysis result.

Use of AI Tools: During the preparation of this thesis, generative AI tools were utilized at several stages, primarily for technical and methodological support. During the methodological development and writing phases: ChatGPT (OpenAI) was utilized to discuss methodological issues, assist with problem-solving, and help with text revision to enhance readability.

Claude (Anthropic) was utilized mostly during data preparation and the analytical stages, helping to design and refine Python code for data extraction, statistical testing, and model validation as well as to generate some figures. Grammarly was used to revise and refine the text focusing on improving its coherence and clarity.

Chapter 1

Literature Review

1.1 Supplier Risk: Foundations and Evolution

1.1.1 Definition and Scope

Within supply chain management, risk is commonly characterized by two dimensions: the probability that a disruptive event will occur, and the magnitude of its impact on the organization should it materialize. A risk attains practical significance only when both conditions are satisfied, namely, when the probability of occurrence and the potential consequences for the firm are each non-trivial [2].

Suppliers occupy a central position within the broader supply network, and mainly this is the reason why that supplier risk holds a comparably central place within supply chain risk management(SCRM). A disruption affecting one supplier generally affects all the chain, reaching production, logistics, and delivery all together. [1].

Supplier risk also is not only about operational concerns but also about financial instability, quality failures, delivery delays, environmental shocks, and geopolitical uncertainty [3]. These factors do not occur in isolation, and this is why a single-dimension view of supplier risk is insufficient. Supplier risk assessment involves identifying these threats, analyzing them, and weighing their implications for supplier performance, with the aim of maintaining supply chain resilience and continuity of operations [1].

1.1.2 Historical Development of Supplier Risk Management

Supplier risk management has changed considerably over the past two decades, mostly driven by the growing complexity of global supply chains. Supplier evaluation was, at that time, mainly an operational and procurement matter, with firms judging suppliers primarily on cost, quality, and delivery performance [2]. Risk management was reactive: a disruption happened when and the firms responded only afterward.

Global sourcing modified supply chain structures in fundamental ways. Suppliers spread across increasingly distant regions, logistics systems extended across national borders, and sourcing relationships grew more intricate than before. This left organizations more exposed to transport disruptions, geopolitical instability, dependency on suppliers that could not easily be substituted, and disruptions originating well outside the firm’s own operations [3].

Sustainability considerations entered the field around this same period. As Sustainable Supply Chain Management (SSCM) took hold, supplier evaluation gradually moved past a purely operational focus. Procurement decisions began to account for environmental compliance, ethical sourcing, labor conditions, and traceability alongside the traditional performance metrics [4].

Several major disruptions have shown how vulnerable interconnected supply networks are. The COVID-19 pandemic, geopolitical conflicts, climate-related events, and regulatory changes have tested traditional supplier assessment methods, revealing their limitations and underscoring the need for approaches that prioritize resilience and account for uncertainty [1, 11]. Christopher and Peck define resilience as a supply chain’s ability to revert to its original state, or to an improved one, after a disturbance cite christopherpeck2004. Later research on supply chain viability which is covered in Section 1.2.2, builds on this foundation.

Recent literature agrees on the same point: the field needs an integrated supplier risk management framework that combines operational performance indicators, external uncertainty factors, sustainability requirements, and data-driven analytical methods [1, 3].

1.1.3 Categories of Supplier Risk

The classification system adopted in this section is grounded in a fundamental separation introduced by Tang [10], who distinguishes between recurrent *operational risks* that emerge from normal uncertainties about supply, demand and costs and disruptive *risks* caused by exceptional events such as natural calamities, strikes, or economic instability. Supplier risks is a generic term that can be divided into different types, which are determined by their source and their effect on the operation of the supply chain. Types of supplier risk are outlined in Table 1.1.

Risk Category	Main Characteristics
Operational Risk [2]	<ul style="list-style-type: none"> • Delivery delays • Quality failures • Production interruptions and capacity shortages • Reliability issues
Financial Risk [1]	<ul style="list-style-type: none"> • Financial instability • Contract fulfillment issues
Geopolitical and Environmental Risk [3]	<ul style="list-style-type: none"> • Political instability • Trade restrictions • Climate disruptions and natural disasters
Sustainability and Compliance Risk [4]	<ul style="list-style-type: none"> • Environmental compliance • Ethical sourcing risks • Labor standard violations • Traceability limitations
Systemic and Network Risk [1]	<ul style="list-style-type: none"> • Cascading disruptions • Supply chain interdependencies • Supplier dependency • Risk amplification

Table 1.1: Main Categories of Supplier Risk

Supplier risks do not exist in isolation. The literature is consistent on this point: such risks are interconnected, and a proper assessment needs integrating operational, financial, sustainability, and external uncertainty indicators into a single, multi-dimensional approach. Lang et al [8] get a comparable conclusion through a recent multi-layered framework, confirming that supplier risk cannot be reduced to one single indicator but must instead be represented by them all together.

1.1.4 Supplier Risk Assessment and Supplier Selection

Risk assessment and supplier selection processes are often used interchangeably in literature. However, these processes are very different from each other. Selection concerns the identification of suppliers that best satisfy an organization's requirements evaluated on different measures like cost, quality, delivery performance and other related criteria. Risk assessment is another process, which involves an analysis of what might go wrong with a given supplier and what consequences would follow for the supply chain.

The importance of this differentiation is that a supplier may satisfy all criteria of selection, but still, represent a risk.

Therefore, risk assessment should not be regarded solely as a prerequisite for supplier selection. It also serves an ongoing role by monitoring engaged suppliers and identifying changes in their risk profiles over time. This ongoing assessment is particularly important in the contemporary context, in which conditions may change significantly after a supplier has been onboarded.

This distinction indicates a gap in the existing literature. Supplier selection has received substantial academic attention, likely because it corresponds to structured performance criteria that are relatively easy to measure and compare. In contrast, risk assessment is more complex, frequently requiring additional data and sometimes sensitive information that is not readily accessible [1]. A framework for supplier risk assessment must address two considerations simultaneously: the methods by which suppliers are evaluated and the availability and dependability of the underlying information.

1.2 Uncertainty in Supply Chains

1.2.1 Sources of Uncertainty

Uncertainty represents one of the primary characteristics of modern supply chains:

- **Geopolitical and Economic Uncertainty:** International conflicts, trade restrictions, political instability, sanctions, inflation, and currency fluctuations can affect supplier continuity, transportation routes, and procurement activities [3].
- **Environmental and Climate Uncertainty:** Natural disasters, floods, extreme weather conditions and climate change can disturb production systems, logistics infrastructures, and agricultural supply chains [3].
- **Regulatory Uncertainty:** New sustainability regulations and traceability requirements impose additional compliance obligations on firms operating in global supply chains [4].

As uncertainty continues to increase, organizations face greater difficulties in identifying and anticipating risks using static and traditional supplier assessment methods.

1.2.2 Supply Chain Vulnerability Under Uncertainty

- **External Disruptions:** Political instability, trade restrictions, currency shifts, natural disasters, transportation failures. None of these need to happen all at once a single one is often enough to interrupt sourcing and put operational continuity in question [3].
- **Ripple Effect Across Supply Networks:** When a supplier or a logistics node goes down, the damage rarely stops there. It tends to show up in other regions, affecting companies that had nothing to do with what caused the original failure [1, 5].
- **Supplier Dependency:** Depending too heavily on one supplier, one raw material, or one sourcing region is a known weak point. Organizations built this way tend to feel a disruption harder and sooner than those that spread their sourcing out.
- **Limited Visibility Under Uncertainty:** Data that arrives late, or incomplete, or scattered across not connected systems, is not inconvenient. It removes the visibility an organization needs precisely when a fast decision is required, and that gap becomes more costly lately under the real uncertainty conditions.

Cost, quality, and delivery performance remain the principal criteria on which most supplier assessments are based, and there is nothing fundamentally wrong with them. The issue as recent literature highlighted multiple times, is that these are static measures being asked to do a job in a dynamic environment. Once risk starts shifting faster than a yearly or quarterly review can track it, cost and quality figures alone stop being enough [1, 2, 5].

That gap is a large part of why supply chain risk management has been shifting toward approaches that are more resilience-oriented and more data-driven, ones that try to bring external uncertainty into the picture from the start instead of reacting to it once something has already gone wrong.

1.3 Supplier Risk Assessment Approaches

1.3.1 Traditional Approaches

Cost, Quality, Delivery reliability for a long time, these were the main dimensions when it came to assessing supplier risk. Audits, scorecards, periodic reviews, expert judgment,

organizations relied on these tools to keep tabs on supplier performance [2].

These methods can be useful for certain conditions. They let organizations monitor suppliers using criteria that were simple enough to apply consistently. But simplicity came at a cost. Traditional approaches tended to be reactive rather than anticipatory, and they leaned mainly on historical performance, telling you what had already happened rather than what might happen next, even as global supply chains grew more dynamic and less predictable underneath them.

One more limitation concerns the extent to which supplier risk assessment depends on managerial decision and subjective judgment [2].

1.3.2 Quantitative Models

As supply chains have grown more complex and more globally distributed, organizations have understandably turned to quantitative methods for supplier risk assessment. These approaches typically integrate operational, financial, logistical, and sustainability-related variables into structured scoring systems, producing different measurable indicators of risk that can be helpful and used in supplier risk assessment [1, 10].

This approach offers substantive value. Turning a qualitative impression, that a supplier appears risky, into something that can be compared and tracked quantitative evidence constitutes a meaningful methodological advance. Multi-criteria models and weighted evaluation systems take this contribution further, enable a systematic comparison between suppliers.

That said, quantitative approaches tend to strain once conditions turn highly uncertain and start shifting quickly, particularly against disruptions that historical data never saw coming and, by their very nature, could not have been built to predict.

1.3.3 Data-Driven and Anomaly Detection Approaches

Advances in data analysis and machine learning have discovered new possibilities for the evaluation and the monitoring of supply risk. Given sufficient operational and external data, data-driven techniques can handle volumes of information beyond the realistic reach of manual analysis, uncovering hidden patterns, detecting anomalies, and identifying emergent dangers before it is fully recognized [1, 5].

Static, traditional evaluation methods struggle to keep pace with this dynamism. Data-driven approaches are by default more adaptive and proactive, and recent literature consistently highlights how the visibility and responsiveness are critical under conditions of uncertainty. Data-oriented approaches are increasingly presented as the means by which organizations achieve this [5].

Anomaly detection warrants particular attention in this context because it avoids a limitation common in much of this literature: the need for a fully labeled dataset of known risk cases, which is often impossible in real organizational settings. Anomaly detection does not depend on such labels; instead, it identifies observations that vary from expected or normal patterns and flags them accordingly.

This capability is especially important for modern supply chains, where uncertainty is common and disruptions in one segment can quickly spread across the entire network [1,5].

1.3.4 Limitations of Existing Methods

Supply chain risk management has never mattered more, and yet the tools available for assessing that risk still fall noticeably short. A number of limitations, in particular, recur throughout the literature:

- **Fragmented Risk Evaluation:** Most approaches still evaluate one risk dimension at a time, rather than adopting the integrated, multi-level perspective that interconnected supply chain risk genuinely requires [1].
- **Reactive Assessment Methods:** Traditional approaches remain, for the most part, backward-looking rather than forward-looking. They rely heavily on historical indicators and reactive management practices, leaving little room to capture uncertainty or change as it unfolds [3,5].
- **Limited Standardization:** Two organizations asked to evaluate the same supplier may well arrive at two quite different conclusions. Subjective judgment and organization-specific methodologies remain common, producing inconsistency that runs across industries rather than within any single one [2].
- **Difficulty Integrating External Factors:** Operational indicators are one thing. Geopolitical risk, sustainability, environmental exposure, and regulatory compliance are different challenges. Most current frameworks still struggle to bring all these factors together in a single assessment.

Taken together, these gaps point to the same conclusion: data-driven approaches need to be integrated into supply risk assessment.

1.3.5 Toward a Universal Supplier Risk Assessment Framework

The limitations of the existing literature raise a fundamental question: can supplier risk assessment be meaningfully standardized across organizations and supply chains that

differ substantially in structure and context? Existing approaches largely do not attempt this. They rely instead on organization-specific criteria and available data sources, so that a framework developed for one industry offers no guarantee of working for another.

Universality, as used in this research, does not mean what the term might initially suggest. It does not imply a single model generating identical risk scores across every supplier, organization, and industry; that was never the intended objective. Rather, universality here refers to a common, structured assessment architecture, one capable of adapting to different contexts while preserving a consistent underlying logic.

Ideally, such a framework would give organizations a shared structure for identifying relevant supplier risks, organizing risk dimensions, integrating different categories of indicators, and applying analytical methods to the data in order to detect potentially risky observations. At the same time, it would need to remain flexible enough to accommodate differences in industry, organizational requirements, and data availability.

A universal framework can be characterized by a small number of defining features:

- **Multi-dimensionality:**

No single indicator is sufficient. The framework must account for operational, financial, logistical, environmental, geopolitical, sustainability-related, and other risk dimensions as relevant to the context.

- **Data integration:**

The framework should be capable of integrating all the different dimensions whenever are available.

- **Adaptability:**

Organizations, industries, suppliers, and available data all differ, often substantially, and the framework needs enough flexibility to accommodate that variation rather than assume it away.

- **Standardization:**

A consistent methodological foundation is needed, one that does not rest on subjective judgment or on whatever a given organization happens to prefer.

- **Data-driven assessment:**

Analytical methods must be able to identify abnormal, potentially risky observations, particularly where a fully labeled dataset of known risk cases simply does not exist.

Whether this framework is truly feasible is still uncertain. Even if it seems strong in theory, it might not work well in practice if the needed information is missing, unavailable, inconsistent, or unreliable. So, we should not judge its universality just by how many risk dimensions or analytical methods it covers. Instead, we need to see if it can handle the limits of real-world data.

This problem is directly linked to ongoing challenges in data availability and reliability, which we discuss in the following sections.

1.4 Risk Indicators and Evaluation Challenges

Assessing supply risk is not only about picking one lens over another. It takes both, external and internal factors together, to actually get at where the operational weaknesses sit when conditions turn uncertain. And that is the direction the literature keeps pointing: organizations, more and more, have to adopt strategies that weave operational performance together with geopolitical, environmental, and sustainability factors, not treat them as separate boxes to check.

Factor Type	Main Indicators	Main Challenges
External Risk Indicators	<ul style="list-style-type: none"> • Political instability • Trade restrictions • Corruption and governance issues • Climate and environmental risks • Commodity volatility • Sustainability regulations 	<ul style="list-style-type: none"> • Heterogeneous data sources [3, 4] • Limited update frequency • Integration complexity • Dynamic uncertainty
Internal Evaluation Indicators	<ul style="list-style-type: none"> • Delivery reliability • Lead times • Quality performance • Shipment consistency • Compliance and traceability • Operational stability 	<ul style="list-style-type: none"> • Subjective evaluation processes • Limited standardization • Reactive assessment methods • Incomplete operational visibility [1, 2]

Table 1.2: External and Internal Risk Indicators in Supply Risk Assessment

The literature shows that effective supply risk assessment increasingly requires integrated and data-driven approaches capable of combining operational indicators with external uncertainty factors.

1.5 Data as a Recurring Challenge

These approaches, despite their methodological differences, converge on a common unaddressed limitation: whether valid and trustworthy data exists to support them in the first place. This issue receives insufficient attention in the literature. It is not a theoretical concern but a practical one, and it frequently shapes which questions are studied, and how, more than the choice of method itself.

Santos et al. [1] address this issue directly in their review of supplier risk research. Obtaining genuine access to company data, they note, is sufficiently difficult that it has shaped the field itself, quietly influencing which topics receive scholarly attention and which do not. Firms are generally willing to share supplier performance figures, provided the supplier remains unnamed, but access to internal production processes or strategic decisions is far more restricted. One consequence follows directly from this pattern: supplier selection has become the most extensively studied topic within the field, in part because it required less sensitive data than other dimensions of supplier risk. A substantial proportion of published models, the authors further observe, rely on simulated rather than real industrial data, precisely because researchers were unable to obtain authentic data. As the authors themselves state, part of the underlying difficulty concerns “the challenges that researchers face in accessing industrial environments” [1].

Duan [3] encounters the same obstacle from a different perspective. The framework proposed in that study is based on a limited number of company case studies, a limitation the author acknowledges explicitly rather than obscuring: “We only looked at a small number of case studies”.

Papa et al. [9] identify a related, though distinct, problem. Their premise is more direct: operational supply chain data is not merely incomplete in certain respects but can be actively unreliable. This constitutes a more difficult problem than missing data, since missing data is at least readily identifiable, whereas unreliable data is not. Values that are present cannot always be taken at face value.

Considered together, these three contributions reveal a single underlying pattern observed from three distinct perspectives. Access to company data remains restricted, which drives researchers toward samples that are small or simulated, and even the data that is obtained cannot always be regarded as fully reliable. Data validity, accordingly, has not constituted a minor issue confined to individual studies. It has instead represented one of the persistent barriers to supplier risk assessment throughout the development of the field.

It follows that the feasibility of a supplier risk assessment framework has never depended solely on the analytical method selected. It depends equally on whether the data

required by that method is available, reliable, and sufficiently consistent for use. This consideration is particularly significant when the objective is a universal framework, since a framework that relies on data that is overly specific, or too rarely available, risks limited applicability beyond the organization for which it was originally developed.

1.6 Research Gap

Current supplier risk assessment methodologies show multiple limitations. Most frameworks prioritize cost, quality, and delivery performance, relying on historical operational indicators while mostly ignoring external uncertainty factors and dynamic disruptions [2]. These evaluations usually look backward rather than forward and often depend on subjective judgment rather than clear, repeatable processes.

Recent studies continue to call for something better: an approach that is more integrated, more data-driven, and better able to improve visibility and adaptability once conditions turn uncertain. Yet almost none of them actually deliver a practical framework capable of bringing internal operational data, logistics information, external risk indicators, and anomaly detection together within a single supplier risk assessment.

The data validity problem raised earlier makes this gap wider. It is not simply that few frameworks combine these information sources; in practice, doing so runs directly into data that is hard to access, inconsistent across sources, or impossible to fully verify. A framework that works only when the underlying data is already clean and complete is of limited use to organizations confronting the kind of fragmented, real-world data researchers keep describing.

A more general question follows: Is it possible to develop a universal supplier risk assessment framework? The literature shows that many ways exist to evaluate individual risk dimensions, but bringing them together into a single, consistent, flexible framework that works across organizations, industries, and data sources remains a challenge.

1.7 Conclusion

The chapter discussed the origins and development of supplier risk management, identified the main sources of uncertainty affecting modern supply chains, and examined the main methods for assessing supplier risk. The analysis shows a clear trend: the discipline has moved from traditional evaluation techniques focused on cost, quality, and delivery performance to more integrated, data-driven approaches.

Alongside this development, a set of persistent challenges has emerged: fragmented risk evaluation, reactive assessment practices, limited standardization, and difficulty in-

tegrating external risk factors. Underlying all of these is a further challenge, limitations tied to data availability and validity, which arguably proves the most persistent of the group.

These observations are the basis for this research. This thesis examines how to integrate foundational internal operational indicators, logistics information, and external uncertainty factors into a single structured framework. It uses different analytical methods to assess potentially risky observations. Data availability and validity are practical challenges to address when assessing whether the proposed framework is feasible.

The following chapter presents the methodology and framework design underlying this research including the analytical approach adopted, as well as the development of the framework itself, in order to address the research question set out in this thesis.

Chapter 2

Research Methodology and Framework Design

2.1 Objectives and Research Questions

2.1.1 Research main Objectives

This research examines how organizations can strengthen supplier risk assessment as supply chain environments grow increasingly uncertain. The central aim is to develop a structured framework capable of combining internal operational indicators, logistics indicators, and external uncertainty factors within one coherent assessment process.

2.1.2 Research Question

Given the limitations listed in current supplier risk assessment methods in Chapter1, this study addresses the following question: How can heterogeneous sources of information be structured and combined to support supplier risk assessment under uncertainty, and to what extent does each information dimension contribute to the identification of potentially risky observations?

To answer this question, the research proposes a data-driven supplier risk assessment framework that combines internal operational indicators, logistics indicators and external uncertainty factors.

2.2 Research Methodology

This research follows the Cross Industry Standard Process for Data Mining (CRISP-DM), originally introduced by Chapman et al. [6] and widely used in data analytics and machine learning projects. It is presented as an iterative process rather than a strictly

linear one, with earlier phases often revisited as understanding of the data improves. That description fits well this research as the framework here was refined more than once as new operational data became available, as Chapter 3 will show in more detail.

The methodology follows six main phases as shown below in the graph [?]:

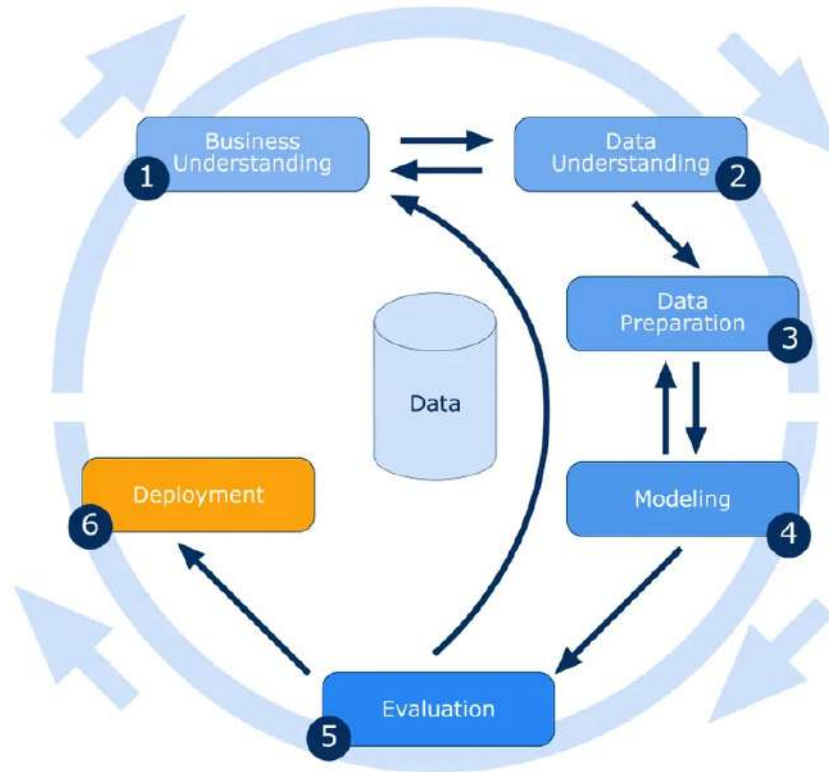


Figure 2.1: CRISP-DM Steps (extracted from <https://medium.com/>).

- **Business Understanding:** Understanding of the supply risk management challenges.
- **Data Understanding:** Exploration of the Different Data sources related to supply chain activities.
- **Data Preparation:** Data extraction, cleaning, preprocessing, integration, and feature engineering processes.
- **Modeling:** Development of complementary data-driven analytical approaches.
- **Evaluation:** Assessment of model performance and interpretation of detected operational risk patterns.
- **Deployment Perspective:** Discussion of potential applications for supply risk monitoring.

2.3 Business Understanding and Proposed Approach

As introduced in the General Introduction, this research was carried out in collaboration with Altromercato, an organization that provided access to internal data on procurement, quality control, and logistics. Altromercato serves as the empirical case through which the broader research problem identified in the literature can be examined against real-world organizational data. Its sourcing model rests on long-term relationships with suppliers rather than short-term contracts, and this distinction carries consequences: a disruption at a supplier does not stop at commercial or logistical effects but can threaten the continuity of the relationship itself or even extending further to other communities that depend economically on that supplier. A sustainability dimension enters supplier risk here precisely for this reason, consistent with the broader discussion of sustainability and supplier risk in Chapter 1.

Beyond this, what makes the case particularly relevant to the research objectives is the availability of data across several organizational processes. It offers a genuine opportunity to examine, in a real-world setting, how heterogeneous information on internal quality, procurement, and logistics can be extracted, structured, and integrated for supplier risk assessment, while also exposing the practical limitations tied to differences in data availability, granularity, and comparability.

Supplier monitoring at this organization spans quality control, logistics, procurement, and shipment tracking, though not every stage of these processes has equal access to information. Some stages are documented constantly, others are only partially recorded, or communicated directly rather than logged through any organized system. This is the main reason behind the data validity issue. The literature already points out problems like restricted access, small or simulated samples, and unreliable data. In this study, these same challenges show up as a mix of formal records, different layout and informal communication that is not recorded.

Traditional supplier evaluation methods, meanwhile, continue to rely on static indicators and manual assessment, which limits their ability to detect emerging anomalies or evolving risks under uncertain conditions. To address these limitations, this research proposes a structured, data-driven supplier risk assessment framework, combining:

- Internal operational indicators extracted from quality control and inspection data
- Logistics indicators related to shipment delays and delivery consistency
- External uncertainty indicators related to geopolitical, environmental, and country-level risks

- An anomaly detection approach used to identify unusual operational behaviors and abnormal supply chain patterns

Together, these dimensions cover the risks directly observable in this study. Quality risk is measured by inspection outcomes, delay risk is assessed using logistics indicators, and geopolitical and environmental risks are evaluated with country-level external indicators. Financial risk was part of the framework in Chapter 1, but because no data are available, it is not included in this analysis.

Rather than combining these three information dimensions from the outset, an approach that might appear more efficient but would obscure what each dimension actually contributes, the framework is instead built step by step, so that each dimension’s contribution can be examined on its own terms before the three are considered together. This owes a clear debt to the multi-layered framework proposed by Lang et al. [8], who similarly argue that supplier risk ought to be constructed from clearly separated information sources rather than collapsed prematurely into a single combined score.

2.4 Data Understanding and Preparation

This study is based on operational supply chain data provided by the company and external factors that are extracted from online sources. The data include procurement, shipment management, logistics activities, delivery monitoring and quality control. These records include information that is related to supplier activity and shipment flows at many stages of the supply chain, including variables that identify suppliers, quality inspection results, transportation and delivery schedules, and country-level contextual information. Each dimension contributes a different perspective to the overall analysis.

The data preparation phase starts by reviewing the recorded documents and understanding each variable. Next, select the relevant documents and extract the necessary data. After that, check for inconsistencies and missing values before preparing the dataset for anomaly detection. The process continues by combining internal operational information with logistics data and external risk indicators. This integration aims to build a complete picture of supplier risk even under uncertainty.

2.4.1 Internal Data

The internal dataset contains operational variables related to supplier quality inspections and shipment monitoring:

- **Supplier and sourcing information:** supplier identity, origin country, raw material category, and product description;

- **Quality control variables:** sampling activities, laboratory process dates, measured quality parameters, and their associated specification limits where available;
- **Inspection outcome:** a recorded Approved or Rejected decision for some inspected shipments, based on client requirements and product quality.

These variables do not originate from a single, uniform source, and that distinction matters from the outset. Inspection records arrive as laboratory reports issued by different quality-control providers, each following its own format, its own layout, its own conventions, and, in many cases, testing an entirely different set of parameters against different thresholds depending on the supplier or the product involved. A report checking moisture content and defect percentages is not directly comparable to one checking heavy-metal contamination, even though both fall under the same category: quality inspections of the same kind of shipment.

The internal data could not, therefore, simply be read from a single table. It first had to be extracted from varied formats, then reorganized carefully, so that inspections testing entirely different parameters could still be examined on some common ground. This extraction and matching process is integral to the internal data dimension, and its execution directly influences the reliability of any resulting indicators. Chapter 3 describes this process, along with its outcomes.

2.4.2 Logistics Data

Logistics data is considered separately from quality data, since delivery performance is a distinct source of supplier risk. This includes:

- Agreed and actual departure dates;
- Document submission dates;
- Expected and actual arrival dates;
- Transit times and delivery consistency across shipments.

There is no specialized logistics dataset available for this study. The only available data is a series of shipment records that include the relevant date details, the expected delivery date, the actual delivery dates, and the departure dates. It was on the basis of these dates, and no other sources, that the logistics variables had to be derived rather than being taken from a specially designed dataset.

Performing calculations across these dates makes it possible to construct variables such as delivery delay and the discrepancy between expected and actual delivery timing. These

derived variables, once constructed, become the basis on which logistics performance is subsequently analyzed. Precisely how they were extracted and calculated, step by step, is presented in Chapter 3.

2.4.3 External Data

External uncertainty indicators associated with supplier operating environments were also considered. These were extracted from the Fairtrade Risk Map(riskmap.fairtrade.net), an external platform that reports country-level risk indicators across several dimensions, including:

- Country-level risk indicators;
- Political and governance information;
- Corruption and regulatory indicators;
- Environmental and climate-related exposure;
- Commodity-related risk information.

These indicators matter because supplier risk does not arise from operational performance alone. It is also shaped by the external conditions surrounding sourcing regions and logistics networks [3]. Their inclusion is intended, specifically, to improve visibility under conditions that are both uncertain and fast-changing.

The Fairtrade Risk Map itself reports these indicators on a categorical scale, ranging from very low to very high risk, rather than as numeric values. This categorical form could not be left as it was. It had to be converted into a form capable of sitting alongside the internal and logistics indicators, so that all three dimensions could eventually be compared, or combined, on the same basis. Precisely how this conversion was defined, and what it implies about the direction and meaning of the resulting scores, is addressed in Chapter 3.

2.4.4 Data Integration

Internal quality records, logistics records, and external country-level indicators are linked using supplier origin and sourcing-related variables, building a fuller picture of supplier risk that combines operational performance, delivery consistency, and external conditions.

Integrating these sources is challenging. The three sources differ in structure, update frequency, completeness, and internal consistency, which is one more concrete version of the same access-and-trust problem raised in Chapter 1 [1, 3, 9]. The difficulty is not that

the sources disagree with each other in some abstract sense, but that getting them to line up took real work. The same supplier, or the same country, was not always recorded the same way across sources: naming conventions, folder structures, and even which field held the country versus the supplier name varied between the internal, logistics, and external data. No single shared identifier matched the records, so harmonizing them required a dedicated step which is described in detail in Chapter 3.

2.4.5 Data Cleaning and Preparation

An important methodological decision concerns how missing values are treated here, and it deserves more attention than it typically receives. A missing value does not always mean lack of value. A variable may simply fail to apply to a particular supplier, product, country, or inspection process: There is no universal standard layout.

For this reason, the cleaning process draws a firm distinction, one carried through consistently rather than applied selectively, between values missing because the underlying information was genuinely unavailable or poorly recorded, and values missing because the variable simply did not apply to that particular observation. This distinction is treated as part of the data preparation methodology itself, not as some inconvenience to be smoothed over with blanket imputation. It holds, without exception, across the internal, logistics, and external data described above.

2.5 Framework Design

The proposed framework brings together internal, logistics, and external information in service of a single aim: supporting supplier risk assessment under conditions that remain, by their nature, uncertain.

Rather than integrating all three information dimensions into a single model from the outset, an approach that would make it difficult, if not impossible, to disentangle what each dimension actually contributes, the framework is instead built incrementally. Internal quality information forms the baseline. From there, additional internal information, logistics information, and finally external information are layered on in turn, deliberately, so that the contribution of each dimension can be examined both on its own terms and in combination with the others. Table 2.1 lists the six scenarios used to test this.

Scenario	Information included
1	Source of Quality 1 (baseline internal quality information)
2	Scenario 1 + additional internal information (Source of Quality 2,3,4)
3	Scenario 1 + logistics information
4	Scenario 1 + additional internal information + logistics information
5	Scenario 1 + external information
6	Scenario 1 + additional internal information + logistics information + external information (all dimensions combined)

Table 2.1: Scenarios used to evaluate each dimension

The purpose of this nested structure is not to determine whether a combined model performs adequately. It is to isolate, with some precision, which dimensions actually contribute useful information and which do not. Two outcomes remain equally possible at the outset, and the evaluation is deliberately built to distinguish between them rather than to assume one in advance.

One possibility is that adding logistics and external information genuinely improves the ability to identify risky observations. A supplier’s delivery behavior, or the conditions prevailing in its country of origin, may well carry information that internal quality data alone does not capture, in which case later scenarios ought to outperform earlier ones. The other possibility runs in the opposite direction entirely: internal information might already be doing most of the work, with logistics and external data contributing little beyond it, or even introducing noise that makes detection harder rather than easier. Under this second scenario, the later models would perform no better, or perhaps worse, than the internal-only baseline.

Neither outcome, it should be emphasized, would constitute a failure of the framework.

2.5.1 Evaluation Strategy

Each scenario is evaluated by comparing model output against shipments whose outcome, Approved or Rejected, is already known. It is precisely this comparison, carried across all six scenarios, that puts the two possibilities described above to a direct test. An improvement from one scenario to the other provides descriptive evidence that the

dimension just added may contain useful information. No improvement, or worse, an outright decline, signals something rather different: that it does not, or at least not in the form and quality available here.

Beyond predictive performance alone, the evaluation also weighs something harder to quantify: the explainability and practical applicability of the framework for organizational supply risk monitoring.

2.6 Conclusion

This chapter presented the methodology and design underlying the proposed supplier risk assessment framework. CRISP-DM guided its development from business understanding through data preparation, modeling, and evaluation, applied here iteratively rather than as a fixed sequence of stages. The framework combines internal, logistics, and external information across six scenarios, using both unsupervised and supervised analysis methods.

Next, we see in chapter 3 how this framework is applied for real-world data.

chapter3

Chapter 3

Data Extraction and Preparation

Chapter 2 defined the three information dimensions underlying the proposed framework: internal, logistics, and external data. Chapter 3 turns to how these sources were extracted, structured, and prepared for empirical analysis, with particular attention to the heterogeneity of the original data, the extraction procedures needed to produce structured observations, and the reconciliation and screening applied to the resulting datasets. These activities correspond primarily to the Data Understanding and Data Preparation phases of the CRISP-DM process described in Section 2.2.

Before proceeding, it is necessary to clarify the level at which supplier risk is observed in the available data. Quality assessments are recorded at the shipment level, not as a direct supplier-level risk label. Each observation is thus a single shipment submitted by a supplier, with an Approved or Rejected outcome reflecting the company’s acceptance decision. Rejected shipments are accordingly treated, throughout this study, as observed instances of supplier-related quality risk.

3.1 Internal Data Extraction

The internal dimension is not a single source. It is made up of four distinct laboratory and document types, Source of Quality 1, Source of Quality 2, Source of Quality 3, and Source of Quality 4, each produced by a different quality-control provider, in a different file format, testing a different set of parameters against different thresholds (as already noted in Section 2.4.1). Following the scenario design in Table 2.1, Source of Quality 1 is treated as the baseline internal source (Scenario 1), while Source of Quality 2, Source of Quality 3, and Source of Quality 4 are treated together as additional internal information (Scenario 2). The two groups differ enough in format and content to need entirely separate extraction logic, so they are documented in two subsections below.

3.1.1 Source of Quality 1 Data Extraction

Source of Quality 1 treated as the baseline internal dataset for that research analysis. The material mainly consisted of shipment-level quality reports stored as PDF documents in year-based company folders covering 2023 to 2026. These documents have shipment information and the laboratory quality measurements at the same time. A dedicated extraction pipeline was therefore developed to convert the reports into a structured, shipment-level dataset.

Document Identification and Extraction

The extraction process first identified the documents relevant to the analysis. Because the company archive contained several document types, classification was performed using characteristic information found in both filenames and document content.

The shipment outcome which is the main label used throughout this thesis to distinguish risky from non-risky shipments was recovered directly from the Source of Quality 1 documents themselves. Each document includes a list of checkboxes at the end. And the checked box indicate the final outcome: Approved, or Rejected or Downgraded. The classification logic was designed to capture every outcome label present in the company archive.

Once the relevant reports had been identified, shipment metadata and quality measurements were extracted from the document text. The resulting variables included the container identifier, country or quality code, shipment date, sample size, and waste percentage, together with the available quality measurements. These included shell content, moisture measured through different methods, pH, cocoa butter, acidity, mold, infestation, insect damage, and germination-related indicators.

The extraction logic required iterative refinement because the reports did not follow a perfectly uniform numerical structure. An initial approach searched for individual numerical values located close to the corresponding parameter label. This proved unreliable when several measurements appeared near one another or when document formatting separated a label from its associated value. The final procedure instead identified the text segment between consecutive parameter labels and extracted the numerical information contained within that segment. This made it possible to accommodate parameters reported with different structures, including minimum and maximum requirements alongside measured values, without assuming that every variable followed the same format.

Documents whose textual content could be read directly from the PDF were processed through the native extraction route. An OCR procedure based on page rasterization and EasyOCR was also developed as an alternative route for scanned or image-based

documents whose text could not be reliably recovered directly. The extraction method used for each record was retained in the resulting dataset. The OCR route formed part of the broader extraction pipeline developed for the company archive, although the 113 Source of Quality 1 shipments ultimately retained for the empirical analysis were successfully extracted through the native text-extraction route. The OCR procedure was therefore not used to generate the final analytical observations, but remained necessary for processing other scanned or image-based document batches encountered during the broader extraction process.

Data Consolidation and Outcome Definition

The yearly outputs were subsequently combined into a single analytical table. During consolidation, dates were standardized, duplicate documents were removed using the source filename, country and outcome labels were harmonized, and the records were ordered chronologically. The resulting Source of Quality 1 dataset contains 113 shipment-level observations described by 29 extracted fields.

Of the 113 shipments, 62 were initially recorded as Approved, 23 as Approved after a press-only downgrade, and 28 as Rejected. For the subsequent statistical and modeling analysis, the press-only downgrade cases were grouped with Approved shipments, producing a final binary outcome of 85 Approved and 28 Rejected observations. This binary classification forms the outcome variable used in the supervised analysis presented in .

3.1.2 Additional Internal Information

Additional internal quality information came from Source of Quality 2, Source of Quality 3, and Source of Quality 4 reports stored in the company’s document archive. These documents were mixed with other material and followed substantially different reporting formats, so a dedicated extraction procedure was required to organize, classify, and standardize the relevant information before evaluating its possible integration with the baseline Source of Quality 1 dataset.

The resulting extraction covered 83 documents, summarized in Table 3.1. Of these documents. The purpose of extracting these additional sources was not to replace Source of Quality 1, but to evaluate whether information generated through other internal quality-control processes could provide additional information beyond that available in the baseline dataset. Their extraction therefore supported Scenario 2 in the scenario design set out in Table 2.1.

Parameter Extraction and Requirement Matching

Each document type was processed according to its own reporting format. Source of Quality 2 organoleptic reports state their own conformity limits directly, so these were read as-is, whereas Source of Quality 2 laboratory reports were used mainly to pull out cadmium, lead, and the laboratory’s own conformity statement. Source of Quality 3 and Source of Quality 4 reports provide physical and sensory quality measurements, including moisture, acidity, moldy, insect-damaged, slaty, violet, and other bean characteristics.

For Source of Quality 3 and Source of Quality 4, the extracted measurements were compared against the company’s technical specifications workbook, which lays out supplier, country, and product-specific requirements. The different specification formats found there were standardized into comparable threshold rules, and supplier and specification matching was done using text similarity. A confidence level was kept for every match, so reliable matches could be told apart from cases still needing manual verification.

The final dataset records the document’s source and type, country, supplier, lot, Outcome, extracted parameters, and the requirement-matching information behind each one. Because the three sources differ substantially in both sample size and parameter coverage, they were not treated as directly comparable at the individual-parameter level, a limitation carried forward into how these data are used later in the analysis.

3.2 Logistics Data Formulation

As already noted in Section 2.4.2, no purpose-built logistics dataset existed. The relevant delay and consistency indicators were instead derived from other variables for different shipments like the actual departure dates, document submission dates, expected arrival dates, and transit times at the shipment level to calculate the variables needed in the analysis. The table below lists the core indicators derived from these dates:

Indicator	Formula / Definition
Departure delay	The difference between the actual departure date agreed and the real departure date (days)
Documentation delay	the period between the actual departure and the Document submission(days)
Documentation status	Based on the documentation delay: Normale (≤ 20 days), In Ritardo (21–40 days), Molto In Ritardo (> 40 days)
Transit time	The period between the expected arrival at port and the actual departure date (days)
Median transit time	Median transit time across all shipments from the same origin country
Arrival stability	Compares a shipment’s transit time to its origin country’s average: "Stabile" if the shipment falls within the normal range for that origin , "Non Stabile" otherwise

Table 3.2: Logistics indicators

The tracker was then restricted to the eight cocoa-sourcing countries relevant to this study : Peru, Ecuador, Togo, Colombia, Madagascar, Uganda, Sierra Leone, and the Dominican Republic leaving 256 shipment-level records.

3.3 External Data Extraction

External uncertainty indicators were obtained from the Fairtrade Risk Map (riskmap.fairtrade.net), introduced in Section 2.4.3. The platform has no public API, and its indicators sit embedded within individual country pages rather than exposed as a downloadable table, so The extraction was performed programmatically using Python. The script first identified the country-page links from the Risk Map’s country index and then retrieved the needed data from each country page recording the corresponding risk category and current performance rating alongside it.

The Fairtrade Risk Map reports risk on five ordinal categories: Very Low, Low, Medium, High, and Very High. These were converted to a normalized 0–1 scale for quantitative analysis, with higher values indicating higher risk, as shown in Table 3.3.

Fairtrade Risk Map rating	Numeric value
Very Low	0.00
Low	0.25
Medium	0.50
High	0.75
Very High	1.00

Table 3.3: Fairtrade Risk Map ratings.

Once converted, the indicator-level data were aggregated to the country level. For each country, an overall `external_risk_score` was calculated as the mean of its available indicator scores, and category-specific risk measures were derived for the dimensions used later in the analysis, including Climate & Forest and Water & Biodiversity. The number of indicators available per country was kept as its own field too, since coverage is not identical from one country to the next.

This produced usable indicators for 120 countries. Only the eight cocoa-sourcing countries relevant to this study were retained for the empirical analysis, linked to the main supplier dataset through the country identifier.



Figure 3.2: Fairtrade Risk Map. Source: <https://riskmap.fairtrade.net/>

3.4 Data Preparation

With the three extraction pipelines complete, the resulting tables were brought together into a common analytical structure. This stage covers cross-source reconciliation, an assessment of how comparable the two internal sources actually are, statistical screening of candidate variables, and the preparation of the final modeling datasets used in Chapter 4.

3.4.1 Structure of the Consolidated Dataset

The consolidated workbook brings together the datasets generated from the Source of Quality 1, internal quality, logistics, warehouse-quality, and external-risk sources. Table 3.4 summarizes the main tables used during data preparation.

Dataset	Rows	Cols	Content
Source of Quality 1 sum	113	16	Summary of Source of Quality 1 records, retaining only the statistically significant variables
Additional internal detailed data (Sources of Quality 2–4)	83	99	Additional quality reports for other shipments, without a recorded Approved/Rejected outcome
Internal summ	83	16	Summary of additional internal records
Logistic	256	16	Shipment-level logistics indicators
External factors	120	13	Country-level Fairtrade risk indicators

Table 3.4: Main datasets list.

3.4.2 Data Quality and Cross-Source Reconciliation

Comparing the sources against each other surfaced inconsistencies in the Source of Quality 1 country and year fields. The country information was recovered from the cleaned Source of Quality 1 sum table using the shared file identifier, and the year was rebuilt from the shipment date instead.

Supplier names also they differed substantially between the Source of Quality 1 and logistics sources, so the logistics dataset was treated as the reference for both supplier and country identifiers. An exact match on country, year, and supplier simultaneously was only available for a small number of shipments, so logistics information was linked through a hierarchical matching strategy instead, trying country-year-supplier first, then country-supplier, then country-year, and finally country alone.

3.4.3 Comparability of Additional Internal Data

Before folding the Source of Quality 2, Source of Quality 3, and Source of Quality 4 records into the modeling dataset, they were checked against Source of Quality 1’s quality variables for comparability. Some Source of Quality 3 and Source of Quality 4 variables do have conceptual equivalents in Source of Quality 1, but Source of Quality 2 measures something else almost entirely, mainly heavy metals and pesticides. Even the variables the sources do share are thinly covered: moisture, for instance, is available for only 8 of the 83 additional internal records.

Merging at the parameter level would therefore have introduced a lot of missingness while shrinking the usable sample considerably. The additional internal data was kept as supporting context instead, left out of the final modeling dataset, which is built from Source of Quality 1’s quality variables combined with logistics and external-risk indicators.

3.4.4 Individual Significance Testing

Candidate variables were first screened individually against the binary outcome, **Approved** versus **Rejected**, using the Mann-Whitney U test. The initial feature-screening threshold was set at $p < 0.05$. Because multiple candidate variables were tested simultaneously, Benjamini-Hochberg false discovery rate (FDR) adjusted values were also calculated and are reported alongside the raw p -values as an additional robustness check, rather than as the criterion actually used to admit a variable into the candidate pool [12].

The screening identified a group of Source of Quality 1 variables showing evidence of association with shipment outcome, including acidity, damaged beans, infested beans, partially violet beans, slaty beans, violet beans, moldy beans, and sample size. Acidity produced the strongest univariate result overall ($p = 0.0003$).

After the logistics and external dimensions were added to the candidate pool, departure delay ($p = 0.0003$) and climate-and-forest risk ($p = 0.0024$) also showed strong evidence of association with rejection. Water-and-biodiversity risk was likewise associated with the outcome in this unadjusted screening ($p = 0.0238$); Section 3.4.5 explains

why it was not carried forward.

Figure 3.3 summarizes the univariate screening across the complete candidate pool, with variables ranked by their unadjusted p -values and colored according to their information dimension. Table 3.5 reports the corresponding sample sizes, unadjusted p -values, effect sizes, and FDR-adjusted values for the variables that cleared the $p < 0.05$ screening threshold.

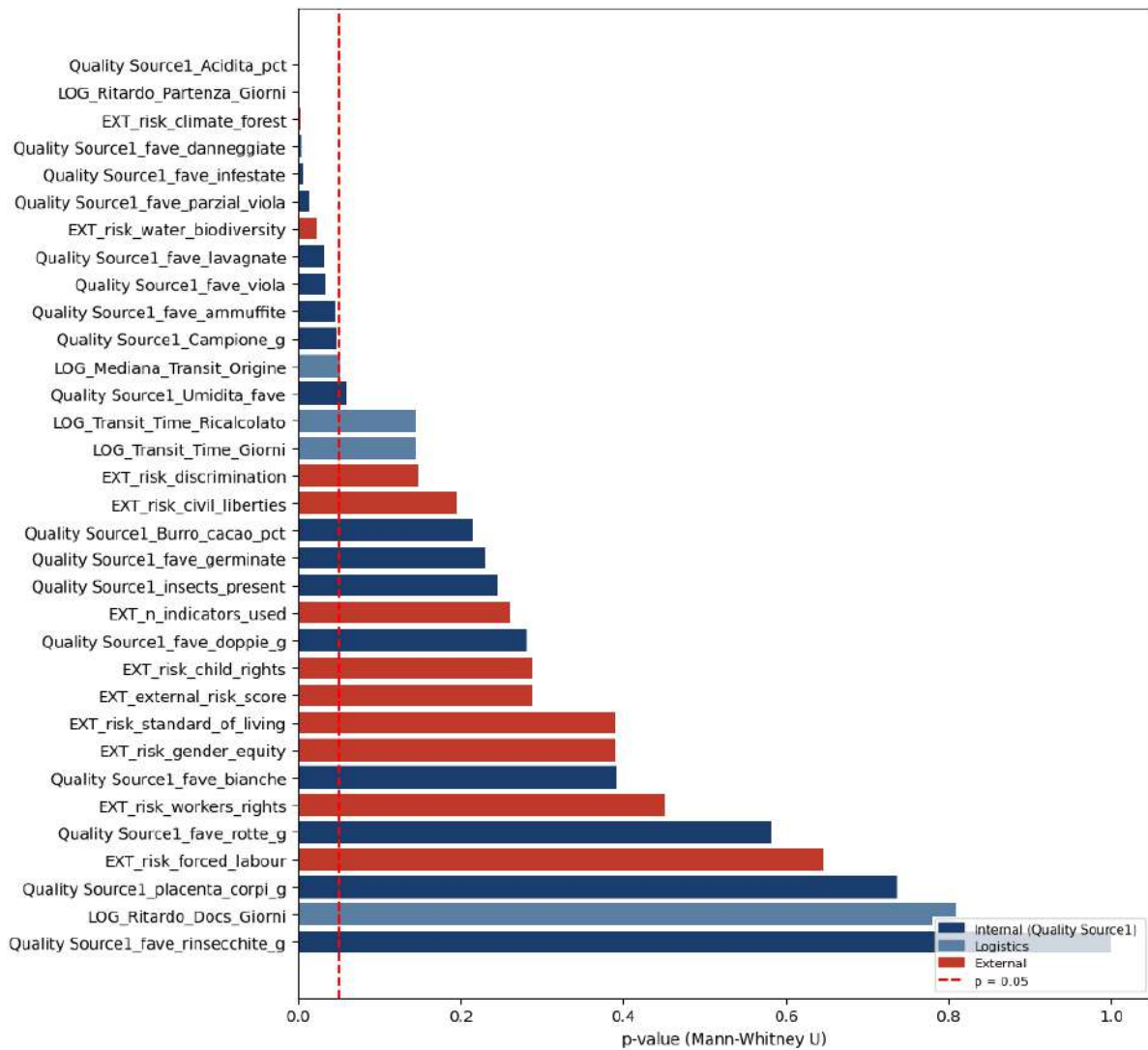


Figure 3.3: Significance test of candidate variables against shipment status

Feature	Source	N	<i>p</i> -value	Effect size	FDR <i>q</i>
Acidity	Source of Quality 1	71	0.0003	0.739	0.009
Departure delay	Logistics	112	0.0003	0.326	0.005
Climate & forest risk	External	112	0.0024	-0.288	0.027
Damaged beans	Source of Quality 1	71	0.0040	0.507	0.033
Infested beans	Source of Quality 1	71	0.0065	0.514	0.043
Partially violet beans	Source of Quality 1	71	0.0139	-0.319	0.076
Water & biodiversity risk	External	112	0.0238	-0.261	0.112
Slaty beans	Source of Quality 1	71	0.0326	-0.210	0.134
Violet beans	Source of Quality 1	71	0.0331	-0.208	0.122
Mouldy beans	Source of Quality 1	71	0.0455	0.442	0.150
Sample size	Source of Quality 1	70	0.0468	-0.272	0.140

Table 3.5: Screening results against shipment status.

Statistical significance depends on sample size and does not describe the importance of an observed relationship alone. Therefore, we reported effect sizes alongside *p*-values. For the binary shipment outcome and the continuous candidate variables, we used point-biserial correlation to measure the association. The *p*-value shows the statistical evidence against the null hypothesis of no association. The point-biserial coefficient indicates both the strength and direction of the observed association. For instance, acidity had a strong positive association with shipment outcome ($r = 0.739$), whereas climate-and-forest risk had a weaker negative association ($r = -0.288$). The associations described here are those observed in the sample and must not be seen as causal effects.

3.4.5 Correlation Screening and Redundancy Reduction

The eleven variables that cleared univariate screening were then examined for pairwise Pearson correlation, in order to identify any pair carrying substantially overlapping information before the candidate set was passed to the regression model.

The strongest relationship by a considerable margin was found between climate-and-forest risk and water-and-biodiversity risk, at $r = 0.94$. As both variables represent closely related country-level risk information, retaining both would have introduced the same external signal into the model twice under two different names. Water-and-biodiversity risk was therefore excluded, and climate-and-forest risk retained as the representative external feature, consistent with the univariate result reported above. Departure delay’s correlation with the Source of Quality 1 sample-size variable, by contrast, was comparatively minor at $r = 0.46$ and did not warrant removal on its own.

The remaining candidate variables exhibited lower pairwise correlations. Several

Source of Quality 1 variables were moderately associated with one another, which is unsurprising given that they describe related physical and quality characteristics of the same shipment (for instance, acidity and damaged beans, $r = 0.66$; acidity and infested beans, $r = 0.58$), yet no other pair approached the degree of redundancy observed between the two external indicators. Figure 3.4 describes the correlation structure of the resulting final feature set.

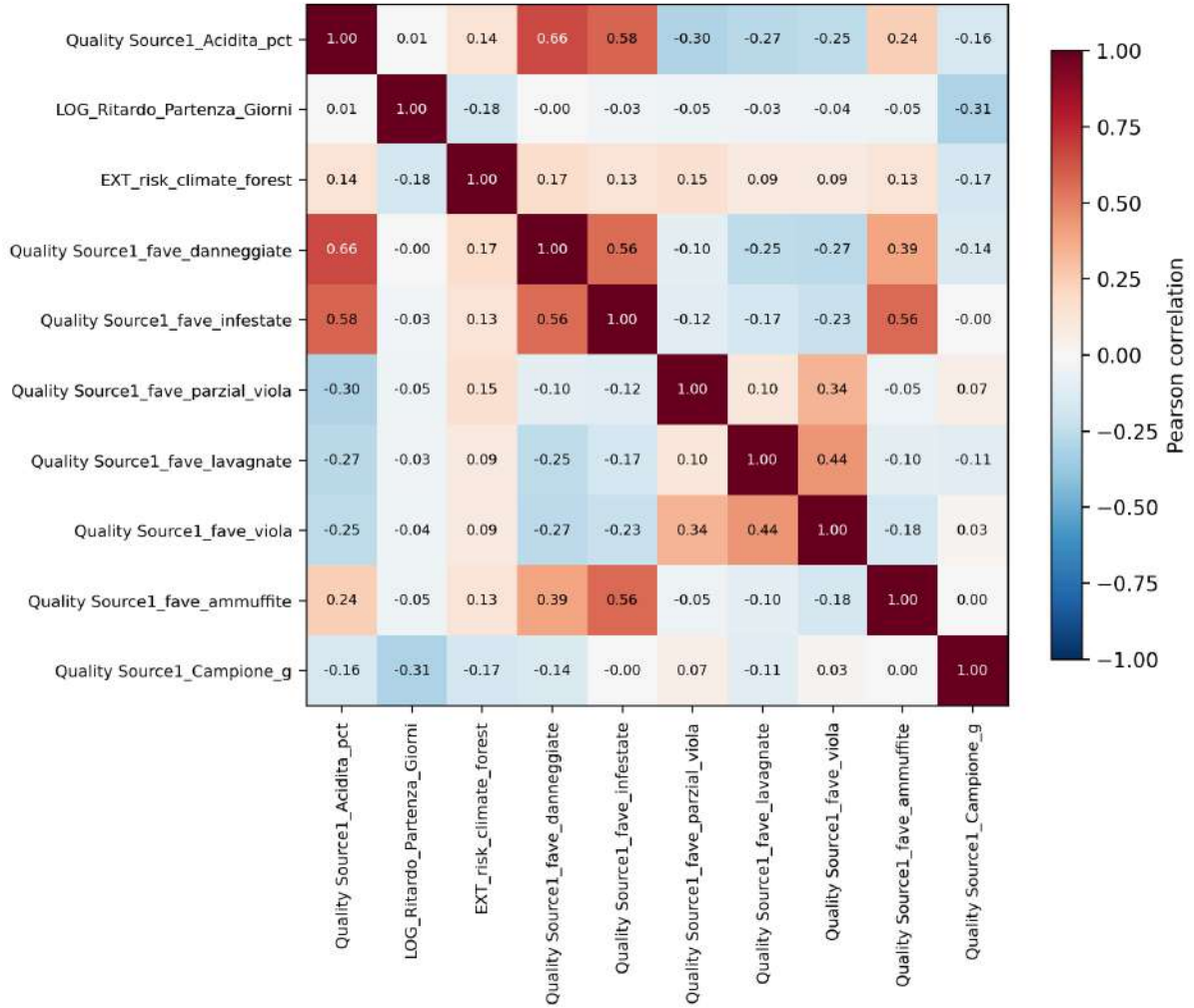


Figure 3.4: Correlation heatmap of the final regression feature set

3.4.6 Multicollinearity Check

As a further check before model estimation, variance inflation factors (VIFs) were calculated for the final ten-variable regression feature set to assess the extent of multicollinearity among the predictors [13]. All VIF values remained below 2.5, indicating that the feature set did not show severe multicollinearity after the redundancy screening in 3.4.5. The highest value was observed for acidity ($VIF = 2.49$), followed by infested beans ($VIF = 2.22$) and damaged beans ($VIF = 2.15$); every other variable in the final set had a lower VIF. This is consistent with, and gives empirical weight to, the moderate correlations already visible in Figure 3.4: the internal defect variables share some variance, but not enough to create a serious multicollinearity problem.

3.4.7 Feature Engineering and Modeling Dataset

These ten variables were carried forward to the modeling stage. Given the small sample size and how much information is missing across several Source of Quality 1 quality fields, the missing-data treatment needed to be decided explicitly rather than left to a default.

Two versions of the modeling dataset were prepared as a result. The first uses listwise deletion, dropping observations without enough available information. The second keeps the full sample and fills missing numeric values with the median of the corresponding variable. Both carry into Chapter 4 for comparison, so the results can be checked for how sensitive they are to this choice rather than resting on a single missing-data decision.

These datasets feed into the CRISP-DM Modeling phase. Chapter 4 evaluates the supervised logistic-regression approach and the six Isolation Forest scenarios defined in Chapter 2 against them.

3.5 Conclusion

This chapter documented how each of the three information dimensions defined in Chapter 2 was turned from raw, heterogeneous source material into a structured table: the baseline data extracted from Source Quality1 bills through an OCR-based pipeline and the additional internal data which is extracted from Quality 2/3/4 reports through folder organization followed by content-based classification and specification matching, logistics indicators through derived-date formulas applied to a general-purpose shipment tracker, and external indicators scraped from the Fairtrade Risk Map, and how large and imbalanced each resulting table turned out to be. It then covered what followed once these tables existed side by side: a country/year contamination problem in the Source of Quality 1 data that only surfaced by cross-checking two of its own tabs against each other,

a supplier-matching exercise that succeeded for only a quarter of shipments, a cascading merge strategy built around that limitation, a direct test of whether the additional internal data could be merged into Source of Quality 1 at the parameter level (it could not, in any useful way), and the significance testing that narrowed a larger set of candidate variables down to the ten actually carried into the modeling stage. Chapter 4 takes this prepared dataset into the Modeling phase of CRISP-DM: fitting and evaluating the supervised regression and the six-scenario Isolation Forest anomaly detector, and comparing what each approach does and does not find.

Chapter 4

Modeling and Results

4.1 Introduction

Chapter 3 produced a prepared modeling dataset, constructed mainly from Source of Quality 1's quality variables in combination with logistics and external-risk indicators, ready to enter the CRISP-DM Modeling phase. This chapter investigates supplier risk through three distinct representations of the problem: an initial exploratory composite risk score, an unsupervised Isolation Forest anomaly detector, and a supervised logistic regression model, each approaching the same underlying question from a different analytical angle.

It should be noted that not all of these approaches survived as final analytical models. The modeling process included preliminary, exploratory attempts, some more promising than others at the outset, that were ultimately discarded once their performance failed to justify further development. These attempts are nonetheless reported here, and deliberately so, as they shaped the development of the final framework in ways that would otherwise remain invisible, and because the rationale for setting them aside is, in its own right, worth documenting.

4.2 Initial Exploratory Composite Risk Scoring

The entropy-weighting method is based on the objective weighting principle used in supplier evaluation literature; it uses entropy to reflect the information or variation in each criterion [14]. Specifically, Mavi et al. [14] combine the Shannon entropy weighting method with fuzzy TOPSIS when selecting suppliers. The present study does not complete full fuzzy TOPSIS. Instead, it uses the entropy component as an exploratory measure to assign objective weights to the available quality indicators and establish a composite risk score. This approach was selected to avoid introducing subjective expert

weights when the available observations and criteria are diverse. The resulting score is therefore an exploratory way to represent the information at hand, not a supplier-selection or decision-making index.

4.2.1 Entropy Weight Method

An initial exploratory approach was developed to test whether the available internal quality indicators could be combined into a single composite risk score. The goal was to turn the different quality deviations observed for each shipment into one common risk representation, without relying on subjective, expert-assigned weights [14].

For this, the Entropy Weight Method was applied to a set of 13 quality-deviation indicators, drawn from the early version of the dataset described in Section 4.3.1. The method assigns higher weights to variables that vary more across the observed shipments, and lower weights to variables that barely differentiate between them; the logic is that a variable which stays roughly the same from one shipment to the next can not be doing much work in telling risky shipments apart from normal ones. Those weights were then applied to the normalized deviation indicators to produce a single composite risk score for each shipment.

The largest weights were assigned to `dev_Moisture_Termo` (0.120), `dev_Moisture_GAC` (0.109), and `dev_Shell` (0.106). This approach was treated as exploratory because it rests on a specific assumption, that heterogeneous quality deviations can be meaningfully aggregated into a single numerical representation of shipment risk, and that assumption needed to be tested directly rather than taken for granted.

4.2.2 Evaluation Against Observed Outcome

The composite score was evaluated against the observed Approved/Rejected outcome using ROC-AUC, rather than simply assumed to represent risk because it was built to. The resulting AUC came out at 0.446, below the 0.5 chance benchmark, meaning the entropy-weighted score did not meaningfully separate Rejected shipments from Approved ones. The mean normalized risk score was also counter-intuitive: Approved shipments averaged 0.382, against just 0.001 for Rejected shipments the reverse of what a working risk score should look like.

That does not necessarily mean the underlying quality indicators have nothing to do with shipment outcomes; several of them, as Section 3.4.4 in Chapter 3 shows, are individually associated with rejection. What it suggests is that, in this dataset, aggregating their deviations into a single entropy-weighted score did not produce a reliable read on rejection risk. This composite-score approach was accordingly dropped as the primary

modeling strategy, and the analyses that follow instead look at the contribution of individual variables and information dimensions directly, either through anomaly detection on the raw feature set or through supervised regression.

4.3 Unsupervised Modeling: Isolation Forest

The Isolation Forest itself follows Liu, Ting, and Zhou’s original algorithm [7], adapted to this setting the way Papa et al. [9] adapt it for supply-chain anomaly detection: rather than training on a mix of normal and anomalous cases, the model is trained only on the population of Approved shipments, treated as the normal baseline, and Rejected shipments are held out entirely from training and used only to check whether the model’s anomaly scores actually separate them from the rest.

This analysis went through three stages before settling into the form reported here.

4.3.1 Initial Exploration

The first Isolation Forest run was performed on the same dataset used for the entropy-weighted composite score in Section 4.2.1. Its ROC-AUC was **0.393**, also below chance, and consistent with the entropy score’s failure to discriminate on the same data. With so few shipments this was treated as inconclusive rather than a real finding at the time, but it set the expectation confirming that this dataset would not make anomaly detection easy.

4.3.2 Scenario-Based Analysis

The analysis was later expanded once the dataset extracted to be aggregated to country-level context and merged in as the additional-internal dimension. This version fed the six scenarios from Table 2.1 directly with raw and spec-deviation features, still training only on the Approved population.

Figure 4.1 shows the result. ROC-AUC moves from 0.253 for Source of Quality 1 alone, through 0.345 once additional internal information is added, 0.378 once logistics is added instead, 0.480 once internal and logistics are combined, 0.489 with external indicators added, and 0.511 once every dimension is combined.

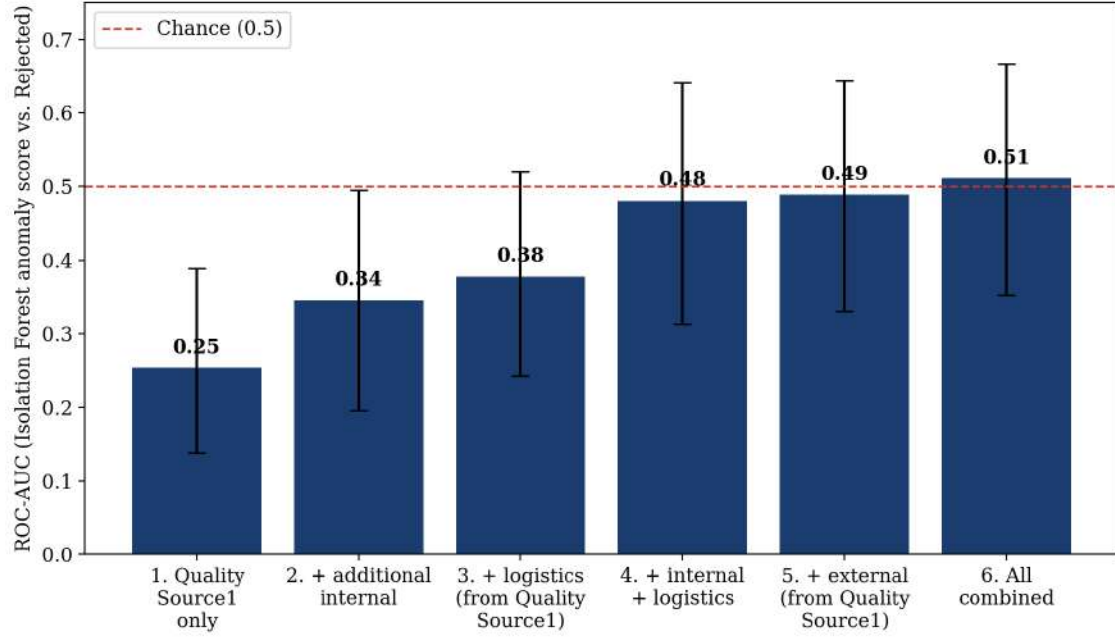


Figure 4.1: Isolation Forest ROC-AUC across scenarios

Table 4.1 gives the exact bootstrap figures behind Figure 4.1: the observed AUC, the bootstrap mean, and the 95% confidence interval for each of the six scenarios (2,000 resamples).

Scenario	ROC-AUC	Bootstrap Mean	CI 2.5%	CI 97.5%
1. Source of Quality 1 only	0.2532	0.2549	0.1386	0.3886
2. + additional internal	0.3447	0.3464	0.1952	0.4948
3. + logistics	0.3775	0.3781	0.2423	0.5200
4. + internal + logistics	0.4805	0.4805	0.3136	0.6410
5. + external	0.4889	0.4893	0.3307	0.6436
6. All combined	0.5111	0.5111	0.3522	0.6667

Table 4.1: Results for the six-scenario Isolation Forest.

This pattern has to be carefully interpreted. The results are not sufficient to support the claim that the Isolation Forest is a dependable rejection detector because the AUC values are still quite low across all six contexts while the bootstrap intervals of the confidence are fairly large as an indication of the limited number of observations available for the test, especially the very small number of Rejected shipments. Still, the sequence of scenarios reveals a slow but steady increase in quality by disclosing more information dimensions. Even this development, however, is not enough for us to argue about the predictability of our model, rather, it suggests a hint that combining different dimensions of information may provide a bigger picture than using only one aspect of the informa-

tion. Making reference to the Table 4.1, only the Scenario 6 produces results higher than what is expected by chance if random classification were used. The direction of the results, nevertheless, is aligned with the conceptual framework laid out in the Chapter 2. More precisely, the results imply that the concept of supplier risk might involve factors like internal product Quality, logistics factors and external-context dimensions, which, altogether explain better than the single dimension that captures them in the isolation of one aspect.

4.3.3 Final Refinement

The third and final version rebuilt the same six scenarios with the internal dimension restricted to Source of Quality 1's eight significant variables (Section 3.4.4), with the additional internal data left out entirely, since Chapter 3 had already established that it could not be merged at the parameter level without producing more missing cells than actual information.

This produced ROC-AUC values between 0.25 and 0.51 showing approximately the same progression pattern as with baseline internal data. This was partly expected, since the additional Source of Quality 2/3/4 records extracted are much fewer compared to the baseline sample.

The remaining scenarios therefore represent a more restricted version of the original design, focused primarily on the significant Source of Quality 1 variables combined with logistics and external information.

Although no refined scenario produced reliable anomaly discrimination, the results remained useful for understanding the limits of the available data. Restricting the feature space to statistically significant variables did not substantially improve Isolation Forest performance, suggesting that the main limitation was not simply the inclusion of weak features. Rather, the combination of a small sample, a limited number of Rejected observations, heterogeneous information sources, and incomplete parameter-level alignment appears to constrain the ability of an unsupervised anomaly detector to learn a stable representation of abnormal shipments.

4.3.4 Leakage-Safe Sensitivity Analysis

As a further robustness check, a stratified 70/30 train/test split was drawn first, univariate significance was recalculated using only the training rows, and only the resulting significant features were carried into an Isolation Forest fitted and scored strictly according to the training/test split. Figure 4.2 report the outcome.

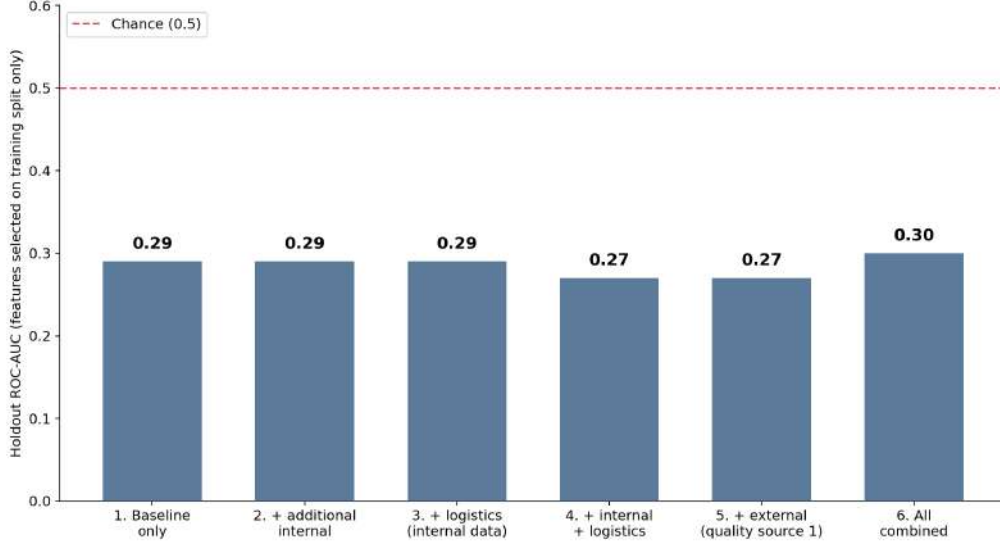


Figure 4.2: sensitivity check by scenario 70/30 split

Even that the leakage-safe analysis produces lower holdout AUC values across all scenarios, it still represents a robust check rather than a replacement for the scenario analysis. Its purpose was to test whether feature selection performed on the full dataset had artificially influenced the evaluation. The lower holdout results indicate that this is not the source of a hidden strong Isolation Forest signal: under a stricter train/test protocol, the available sample remains insufficient for reliable anomaly discrimination.

Taken together, the exploratory scenario progression and the leakage-safe evaluation lead to a balanced conclusion. The different information dimensions appear to contribute complementary information to the representation of supplier risk, but the current volume and structure of the data do not allow Isolation Forest to transform that information into a stable generalizable detector.

4.4 Supervised Modeling: Logistic Regression

Logistic regression models the probability of a binary outcome, here Rejected versus It has been approved on the basis of one or more predictor variables. The model provides an estimate of a set of coefficients that indicate how each predictor is associated with the probability of a shipment being labeled as Rejected.

The logistic regression model can be described as:

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} \quad (4.1)$$

where p_i represents the probability that shipment i is classified as Rejected, β_0 is the

intercept, β_j represents the coefficient related to the predictor X_j , and k is the total number of predictors that are included in the model.

The coefficients can also be expressed in terms of odds ratios as:

$$OR_j = e^{\beta_j} \quad (4.2)$$

4.4.1 Model Specification and Missing-Data Treatment

Logistic regression models the probability of a binary outcome, here Rejected versus Approved, as a function of one or more predictor variables, fitting a set of coefficients that show how each predictor pushes that probability up or down. Logistic regression was used as the supervised component of the analysis, with shipment status treated as a binary outcome distinguishing Approved and Rejected shipments unlike the entropy-weighted score and Isolation Forest approach that do not use the observed outcome when training the model [15].

With 113 shipments, only 28 of them were rejected. That limited sample size and number of observed rejection outcomes were one of the most important considerations in model selection. That is why a relatively simple and interpretable logistic regression was selected over other more flexible modeling approaches that would require stronger empirical support from a substantially larger sample.

The missing-data question raised in Chapter 3 mattered enough to be handled two ways rather than one. Several of the Source of Quality 1 cut-test fields are populated for only a subset of shipments, and that subset does not fall across Approved and Rejected cases, so how to treat missing values was not a neutral choice. A list-wise-deletion version of the dataset, keeping only shipments with enough of the ten variables filled in, was fit alongside a second version that kept the full 113 shipments and filled gaps with each variable's median.

4.4.2 Regression Results

The median-imputed version, keeping the full 113 shipments, is the one worth reporting. This model reached a pseudo- R^2 of 0.272, a moderate fit for a 113-row dataset. Table 4.2 reports the full set of the ten coefficients:

Variable	Coefficient	Std. Error	Odds ratio	<i>p</i> -value
Departure delay	+0.086	0.030	1.09	0.0043
Climate & forest risk	−8.115	4.025	0.0003	0.0438
Acidity	+2.180	1.223	8.84	0.0746
Sample size	+0.00014	0.00015	1.00	0.3221
Partially violet beans	−0.079	0.095	0.92	0.4040
Slaty beans	−0.088	0.129	0.92	0.4934
Damaged beans	+0.037	0.131	1.04	0.7794
Violet beans	−0.055	0.178	0.95	0.7556
Infested beans	−0.070	0.269	0.93	0.7946
Mouldy beans	−0.005	0.164	0.99	0.9739

Table 4.2: Logistic regression coefficients (N=113, 28 Rejected)

Two of the ten variables reach conventional significance in this model: logistics departure delay ($p = 0.0043$) and external climate-and-forest risk ($p = 0.0438$), both already flagged as candidates in Chapter 3 univariate screening. Acidity, despite producing by far the strongest univariate result in Section 3.4.4 ($p = 0.0003$), falls just short of the conventional threshold here ($p = 0.0746$) once the other nine variables are controlled for; Chapter 5 returns to why a variable can behave this way. The remaining seven Source of Quality 1 variables and sample size do not reach significance once acidity and the other defect measures are in the same equation, consistent with several of them describing overlapping aspects of the same underlying spoilage problem.

The broader picture is that the predictive signal in this model is not confined to a single information dimension. The ten variables carried into the regression span all three dimensions proposed in the framework, internal quality, logistics, and external context, and the two variables that reach significance come from two different dimensions rather than both from the same source. Acidity’s near-significant coefficient keeps the internal quality dimension in the picture alongside logistics and external risk, rather than the regression collapsing onto a single variable or a single dimension.

4.4.3 Holdout Evaluation

Evaluated on a held-out 30% stratified split of the data (test $N = 34$, 8 Rejected), this model reached a ROC-AUC of 0.721. Figure 4.3 shows the corresponding ROC curve, and Figure 4.4 shows the confusion matrix at the default 0.5 probability threshold: 14 true negatives, 12 false positives, 1 false negative, and 7 true positives. The high false-positive count at this threshold reflects the class weight=’balanced’ setting used to compensate for the training set’s class imbalance, which trades some precision on Approved shipments for

substantially better recall on the minority Rejected class (7 of 8 Rejected test shipments correctly flagged). With only 34 shipments in the test set, this AUC should be read as a useful signal rather than a precise, stable estimate; a different random split would very plausibly move it by several hundredths.

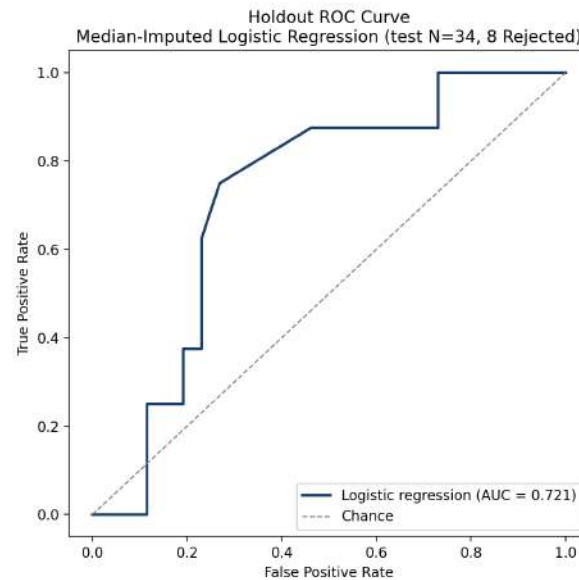


Figure 4.3: Holdout ROC curve, median-imputed logistic regression

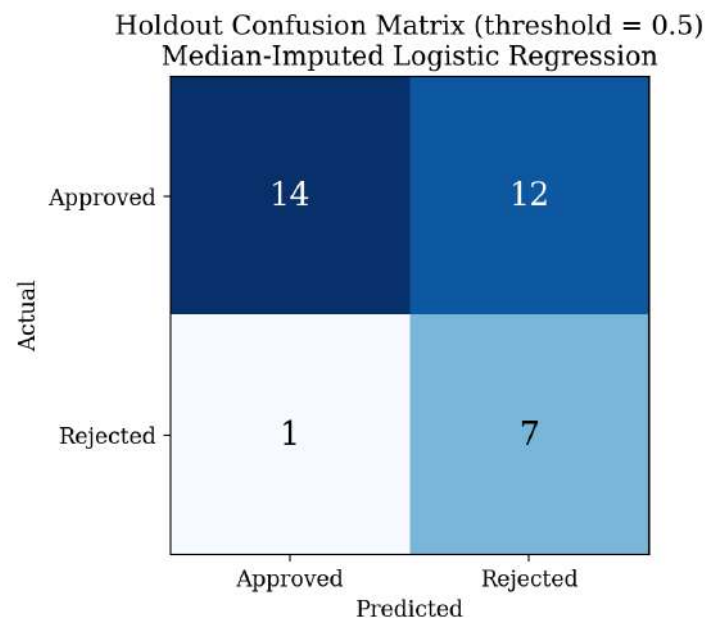


Figure 4.4: Confusion matrix

Table 4.3 decomposes the same test-set predictions differently from the confusion matrix above. Recall (sensitivity) of 87.5% confirms that the model identifies most Rejected

shipments (7 of 8); specificity of 53.8% and precision of 36.8% show the cost of that recall in extra false positives; and the F1-score of 51.9% summarizes the trade-off in a single number.

Metric	Test result
ROC-AUC	0.721
Recall / Sensitivity	87.5%
Specificity	53.8%
Precision	36.8%
F1-score	51.9%

Table 4.3: Holdout classification metrics at the default 0.5 probability threshold.

To check whether the single-split ROC-AUC of 0.721 reflects a stable underlying signal or simply a favorable draw of which eight Rejected shipments landed in the test set, the same model was re-evaluated under repeated stratified 5-fold cross-validation (10 repeats, 50 folds total). This produced a mean ROC-AUC of 0.818, noticeably higher than the single-split estimate. However, given the dispersion across folds and the limited number of Rejected shipments, the cross-validation result should be interpreted as supplementary evidence rather than as a precise or definitive estimate of out-of-sample performance.

4.5 Comparison of Modeling Approaches

Put side by side, the three approaches tested in this chapter do not just disagree on a number; they represent three different ways of asking the same underlying question, and they disagree on whether a usable answer ever exists. Table 4.4 summarizes the comparison.

Approach	Learning type	Result	Interpretation
Entropy-weighted composite score	Composite scoring, unsupervised construction	ROC-AUC = 0.446 Approved mean higher than Rejected mean	The tested aggregation of heterogeneous quality deviations into a single score is not supported by the observed outcomes.
Isolation Forest	Unsupervised anomaly detection	Scenario AUCs increase descriptively from 0.253 to 0.511;	Additional information dimensions appear to enrich the representation of supplier risk, but the current dataset is insufficient for stable unsupervised anomaly discrimination.
Logistic regression	Supervised	ROC-AUC = 0.721	When the observed rejection outcome is used directly, the combined dataset reveals a usable predictive signal involving variables from multiple information dimensions.

Table 4.4: Comparison of the three modeling approaches.

The comparison between the three approaches suggests that the main difference in performance is not simply a question of algorithmic sophistication, but of how supplier risk is represented and what information the method is allowed to use. The entropy-weighted score and Isolation Forest both attempt to identify risk without direct access to the Approved/Rejected outcome. Their results therefore depend on the assumption that rejection can either be represented by a common aggregate deviation or appears as an abnormal pattern relative to Approved shipments.

The scenario analysis nevertheless provides an important exploratory finding. Although Isolation Forest did not achieve reliable discrimination, the progressive increase in AUC observed when internal, logistics, and external information dimensions were introduced is consistent with the idea that these dimensions contain complementary information about supplier risk. The highest exploratory AUC was obtained when all dimensions were combined, suggesting that risk is not fully represented by internal quality information alone.

Logistic regression addresses the same problem differently by using the observed shipment outcome directly. Under this supervised representation, the combined dataset produces a stronger and more interpretable signal. The significant and near-significant predictors are not confined to one information dimension: the results include internal

quality-related variables together with logistics and external-risk indicators. This supports the broader premise of the proposed framework that supplier risk can be influenced by multiple dimensions operating simultaneously.

The contrast between the approaches therefore does not imply that the additional information dimensions were irrelevant to the unsupervised analysis. Rather, the Isolation Forest results suggest that the available sample is too limited for the complementary information contained in those dimensions to be converted into a stable anomaly pattern. When the observed rejection labels are made available through supervised learning, however, the same broader information structure becomes more useful for distinguishing shipments associated with rejection.

4.6 Conclusion

This chapter tested three different ways of representing supplier risk, all built on the dataset prepared in Chapter 3. Rather than reporting only the strongest-performing approach, it documents the exploratory stages that led to the final framework, since each stage shaped what came after it.

The entropy-weighted composite score asked whether heterogeneous quality deviations could be collapsed into a single risk number. They could not: the resulting ROC-AUC of 0.446 fell below chance, and the score’s relationship with the observed Approved/Rejected outcome ran in the wrong direction. For this dataset, entropy-based aggregation simply did not capture the rejection process in one composite value.

Isolation Forest examined the problem from a different angle by testing whether Rejected shipments could be identified as anomalous relative to the population of Approved shipments. The model did not achieve stable or reliable discrimination, and the leakage-safe sensitivity analysis confirmed that its out-of-sample performance remained weak. However, the scenario-based analysis produced a descriptive progression as additional information dimensions were introduced. Performance increased from Source of Quality 1 alone to progressively richer combinations of internal, logistics, and external information, with the highest exploratory AUC obtained when all available dimensions were combined.

The supervised logistic regression performed better. Using the observed Approved/Rejected outcome directly, the model reached a holdout ROC-AUC of 0.721. Its signal was not confined to one source: the ten candidate variables carried into the regression span internal quality, logistics, and external risk, and significance was not limited to a single one of those dimensions.

Taken together, the three approaches tell a more informative story than a simple pass/fail comparison between models. The entropy-weighted score could not reliably

compress the available quality deviations into one common risk number. Isolation Forest pointed toward the same conclusion from another angle: additional information dimensions appear to contribute to the representation of supplier risk, even though the current sample is too small for that contribution to form a stable unsupervised anomaly pattern. Only when the observed outcome was incorporated directly, through logistic regression, did a more usable predictive signal emerge across the combined information structure.

The findings support a multi-dimensional view of supplier risk, but they also point to a methodological limitation: with only a limited number of observations and rejection events available, the value of combining internal, logistics, and external information may be easier to detect through supervised modeling than through the unsupervised approaches tested here. Chapter 5 discusses what these findings mean , and this work can achieve more in the future.

Chapter 5

Discussion

This chapter analyzes the results presented in Chapter 3 and Chapter 4 in relation to the research question and the literature reviewed in Chapter 1. The study examined how internal, logistics, and external information can be structured and combined to support supplier risk assessment under uncertainty, and the extent to which each information dimension helps identify potentially risky shipments.

The discussion therefore considers not only the performance of the individual analytical models, but also what the results reveal about the practical challenges of integrating heterogeneous information for supplier risk assessment. The findings are examined in relation to existing research, followed by their practical implications for Altromercato, the limitations of the study, and directions for future research.

5.1 Theoretical Implications

5.1.1 Answering the research question

We see from the results that all three information dimensions provide relevant signals for potentially risky shipments but they do that in different ways.

Each of the sources (Internal, External, Logistic) adds a specific information. The impact of the different sources depends on the model used and data provided.

Uni-variate analysis investigates to what extent an individual variable is able to differentiate Approved and Rejected shipments without taking into account any other variables whereas multivariate regression looks into whether a variable adds to an already built model by delivering new information. This can be that even if a variable is highly correlated with the label rejection the impact can go downwards in case it is positively correlated.

The acidity variable approved this point: Despite having the strongest univariate

connection to rejection among the variables, its predictive ability gets lowered in the multivariate model.

With that distinction, the pattern across the three information dimensions is as follows. Acidity produced the clearest univariate result of any variable tested ($p = 0.0003$, Table 3.5), making internal quality data the strongest single signal in the dataset at the univariate level. Once the other nine variables are added into the regression, however, it is logistics departure delay that keeps a stable, statistically significant positive association with rejection ($p = 0.0043$), with external climate-and-forest risk as the second significant multivariate term. External country-risk indicators are the weakest of the three dimensions taken in isolation, and the Isolation Forest scenario comparison in Figure 4.1 does not establish that adding them, or logistics, reliably improves detection in an absolute sense; what it shows is a consistent direction of movement as information is added, which is a more modest claim. Taken together, no single dimension is redundant here, but which dimension looks most important depends on whether the question is univariate association, multivariate contribution, or anomaly-detection performance, and the three do not point to the same answer.

5.1.2 Interpretation of the Regression Coefficients

Table 5.1 reproduces the ten-variable logistic regression that Chapter 4 introduced, fit on the median-imputed dataset (N=113, 28 Rejected, pseudo- $R^2=0.272$).

Variable	Coefficient	Odds ratio	p -value
Departure delay	+0.086	1.09	0.0043
Climate & forest risk	−8.115	0.0003	0.0438
Acidity	+2.180	8.84	0.0746
Sample size	+0.00014	1.00	0.3221
Partially violet beans	−0.079	0.92	0.4040
Slaty beans	−0.088	0.92	0.4934
Damaged beans	+0.037	1.04	0.7794
Violet beans	−0.055	0.95	0.7556
Infested beans	−0.070	0.93	0.7946
Mouldy beans	−0.005	0.99	0.9739

Table 5.1: Logistic regression coefficients.

Departure delay is the cleanest result in this model: a positive coefficient, significant at $p = 0.0043$ even with nine other variables in the equation, in the direction the framework anticipated in Chapter 2. Unlike the internal defect variables, departure delay

does not share much variance with any of them, so its effect is not diluted by overlapping predictors the way acidity's is.

Acidity keeps a positive coefficient, consistent with its univariate result, but falls just short of the conventional 0.05 threshold once the other nine variables are in the model ($p = 0.0746$). Acidity correlates at 0.66 with damaged beans and at 0.58 with infested beans in this sample, so once those two sit in the same equation, some of acidity's explanatory power gets shared out rather than lost. The most defensible statement is that acidity has the strongest univariate relationship with rejection of any variable in this dataset, and the multivariate model's weaker significance reflects overlap with correlated defect variables rather than a weaker underlying effect.

Climate-and-forest risk reaches significance in the multivariate model ($p = 0.0438$), but its sign is the most counter-intuitive result in this thesis: higher country-level climate risk is associated with lower odds of rejection. A more likely reading is that this coefficient is picking up something about which countries happen to supply which document types, rather than a real causal relationship between climate exposure and shipment quality. This is discussed further as a limitation in Section 5.3, and it is not a result this thesis is willing to stand behind as evidence that climate risk lowers rejection risk in any real sense.

The remaining seven variables all fail to reach significance once acidity and departure delay are controlled for. Since several of these visibly correlate with acidity or with each other, describing different facets of the same underlying spoilage problem, the most reasonable interpretation is that acidity and departure delay jointly capture most of the information these variables also carry.

5.1.3 Comparison with Literature

This thesis applies the isolation forest approach used in **Papa et al.** [9]. The final outputs with the type of data used in that model do not confirm the results reported at **Papa et al.** All the attempts for this model produced poor evaluation metrics scores. This points toward a mismatch between what Isolation Forest looks for (a shipment that is a multivariate outlier across many dimensions at once) and what actually exists in the real data.

Tang's [10] This distinction between recurrent operational risk and rarer disruption risk maps onto what was modeled here. Acidity and departure delay are operational risk indicators in the sense of Tang. The external Fairtrade indicators were constructed to capture disruption risk, but their poor or counter-intuitive performance indicates that, at the present scale of this framework, it should be treated as an operational risk tool rather than a disruption risk tool.

Christopher and Peck's [11] notion of resilience is a different construct from anything measured here. A shipment-level rejection model says nothing about Altromercato's capacity to absorb and recover from a chain-level disruption, which is closer to what Ivanov et al. [5] describe as supply chain viability. This framework is best read as one possible input into a resilience assessment, not a resilience metric on its own.

Lang et al.'s [8] argument that supplier risk should be built from clearly separated information sources rather than a single combined score is reinforced here: the two modeling approaches disagreed on whether a usable signal ever existed, a disagreement a single composite score would have hidden.

Santos et al. [1], **Duan** [3], and **Papa et al.** [9] flagged data validity as a persistent, under-discussed barrier in this field. Chapter 3 turned that general claim into a fully worked example: the contaminated Country and Year fields, the supplier-matching exercise that succeeded for barely a quarter of shipments, and the comparability test that ruled out merging two internal sources are a concrete demonstration of exactly the access-and-trust problem these papers describe from a distance.

5.2 Practical and Managerial Implications

For Altromercato specifically, the most immediate value of this research is not the model, whose performance does not yet support using it for real approve or reject decisions, but the structural picture it exposes of the three information axes the framework is built on. Figure 5.1 explains the current status of each axis and what could be better for a future version of this framework to run on more solid ground.

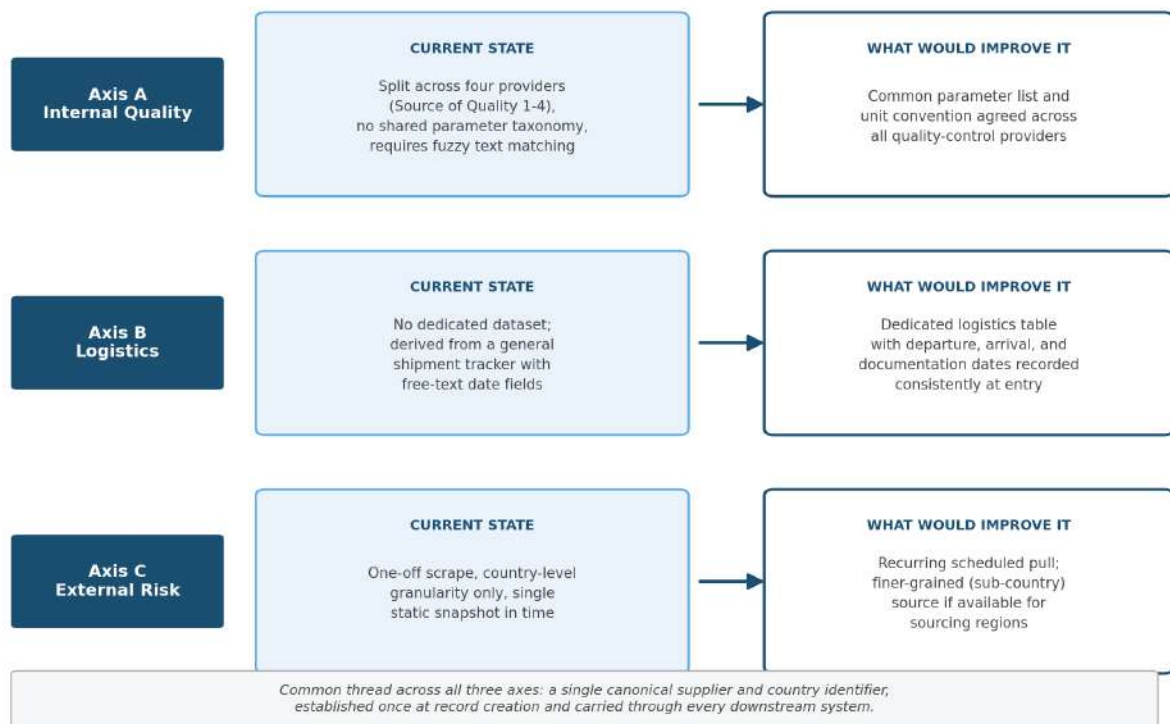


Figure 5.1: Current versus desired state of each information axis Generated by Claude (Anthropic)

Axis A, internal quality, is split across four providers with no shared way of describing the same defect. This thesis had to build that shared vocabulary after the fact, through fuzzy text matching and manual specification lookups. Where operationally feasible, greater standardization of the parameters and units reported across quality-control providers would substantially improve future comparability, so that it is built in from the start rather than reconstructed by a researcher months later.

Axis B, logistics, does not exist as a separated dataset. It was derived from a general shipment tracker never designed to answer questions about delay or consistency. A dedicated logistics table would remove a category of parsing work and solve the country/year alignment problem considerably.

Axis C, external risk, currently depends on a one-off scrape run once for this thesis, pinned to country-level granularity. A Time based Data would keep this axis current.

The common thread across all three axes is the absence of a shared supplier and country identifier, established once at the point each record is created rather than reconciled after the fact for each new research question.

Given the current sample size and data limitations, the findings are most useful for

supporting exploratory review, rather than for automated decision making. Variables such as departure delay and acidity may be useful as inputs to prioritize shipments for further assessment, particularly as they demonstrated the strongest evidence of association with rejection across the analyses. However, the present results do not define validated operational thresholds and should not yet be translated to automatic approval, rejection or fixed alert rules without further testing on additional data.

5.3 Limitations

- **Sample size:** The central constraint is sample size: 113 shipments, of which only 28 are rejected, provide a limited base for a ten-variable regression. This directly motivated supervising more flexible models before trying this one, explains the outright failure of the listwise-deletion regression through near-perfect separation once incomplete rows were dropped, and means individual coefficient estimates, particularly acidity's, should be interpreted as indicative of direction rather than precise, stable values.
- **Extraction accuracy:** Extraction accuracy from the underlying OCR and document-parsing pipelines was not independently verified against source documents on a random sample. Some unmeasured extraction error therefore likely remains in the final dataset, notwithstanding the classification and field-extraction refinements described in Chapter 3.
- **Incomplete integration of additional quality-control data:** Additional internal quality data from three of the four quality-control providers could not be incorporated into the final model, for the comparability reasons discussed previously. The analysis in Chapters 4 and 5 consequently draws on a narrower slice of Altromercato's total quality-control effort than exists in principle.
- **Static external indicators:** External risk indicators constitute static, country-level snapshots rather than supplier-specific, time-varying measures. They therefore cannot capture a supplier's improvement or deterioration over time, reflecting only the general risk profile of the country in which it operates.
- **Absence of financial and economic dimensions:** The available data lack sufficiently detailed financial and economic information to capture dimensions such as procurement costs, the costs associated with rejected shipments, replacement or switching costs, contractual penalties, or the broader financial consequences of supplier disruption. This thesis therefore speaks to operational risk detection without

addressing the financial stakes attached to any given prediction.

- **Limited generalizability:** Every result in this thesis derives from one company, four quality-control sources, and eight sourcing countries over a limited time window. It therefore cannot be assumed that they can be applied to other companies, commodities, or supply chains without additional empirical verification.

Chapter 6

Conclusion and Future Research

6.1 Conclusion

This thesis set out to answer a specific research question: how can heterogeneous sources of information be structured and combined to support supplier risk assessment under uncertainty, and to what extent does each information dimension contribute to the identification of potentially risky observations? This research question has two parts: at the data level, it examines which information sources matter, and at the methodological level, it examines which analysis methods can capture and prospect their importance. Chapters three to five discuss both areas together.

On the data side, no single dimension gives a complete picture on its own. Internal quality data carries the strongest individual association with shipment rejection: acidity alone produced the strongest univariate result of any candidate variable, and several other Source of Quality 1 defect measures, damaged beans, infested beans, partially violet beans among them, cleared the same $p < 0.05$ screening threshold. That strength does not survive intact once the variables are placed together in a single model. However, acidity's coefficient in the ten-variable logistic regression falls to $p = 0.0746$, just short of conventional significance, because multiple of the internal defect variables are moderately correlated with each other and therefore contain overlapping information. (Section 3.4.6 put a number on this: acidity's variance inflation factor of 2.49 was the highest in the final feature set, but still well short of the range that would indicate a serious multicollinearity problem). Logistics information, represented here by departure delay, turns out to be the most robust multivariate predictor once all ten candidate variables are placed in the same model, remaining significant at $p = 0.0043$ even after acidity's univariate advantage is controlled for. External country-level risk indicators are the weakest of the three dimensions on their own and one of their coefficients, climate-and-forest risk, runs in a direction that is harder to reconcile with the framework's own working assumptions

than departure delay’s does, yet they still moved Isolation Forest’s anomaly-detection performance in a consistent upward direction when added on top of internal data, from an AUC of 0.253 with Source of Quality 1 alone to 0.511 once every dimension was combined. Contribution, in other words, depends on which question is being asked, whether that question is univariate association, multivariate significance, or anomaly-detection performance, and the three questions do not point to the same variable as the most important one.

A second, equally important part of the answer concerns method rather than data. Three different representations of supplier risk were tested against the same shipments: an entropy-weighted composite score, an unsupervised Isolation Forest anomaly detector, and a supervised logistic regression. The supervised method, because it had direct access to the known Approved/Rejected results, was the only one to meet the standard statistical thresholds and perform well on the held-out sample, achieving a pseudo- R^2 of 0.272 and a holdout ROC-AUC of 0.721. Further cross-validation, carried out 50 times with 10 repeated 5-fold partitions, indicates that the outcome was, if anything, conservative rather than chance-based. The two label-blind approaches, the composite score and the anomaly detector, both failed to establish reliable discrimination, and did so consistently across multiple independent attempts rather than because of a single implementation choice: the entropy-weighted score reached an AUC of only 0.446, actually ranking Approved shipments as riskier than Rejected ones on average, and Isolation Forest’s best six-scenario result of 0.511 sat inside a bootstrap confidence interval wide enough to include chance. This is itself part of the answer to the research question: under the data constraints documented throughout this thesis, structuring and combining information sources is necessary but not sufficient; the analytical approach used to interrogate that combined information matters just as much, and a framework that only tested one method would not have been able to tell the two apart.

6.2 Theoretical and Methodological Contribution

This thesis contributes to the supplier risk literature in four specific ways.

- **A structured comparison of multiple information dimensions:** Rather than assuming from the outset that internal quality, logistics, and external context should be combined into a single model, this thesis tested their contribution incrementally, through the six-scenario design introduced in Chapter 2. This approach made it possible to observe that internal quality data accounts for most of the effect in absolute terms, while logistics and external context contribute smaller, direction-consistent effects, a distinction that a single combined score would have concealed

entirely, since a composite index collapses precisely the scenario-by-scenario comparison that revealed this pattern in the first place. This follows the multi-layered logic proposed by Lang et al. [8], and the results obtained here reinforce their argument that supplier risk is better represented through clearly separated information sources than through a single index, while also providing that argument with an empirical demonstration on a real, if modestly sized, supplier dataset rather than a purely conceptual one.

- **A structured comparison of composite scoring, supervised learning, and anomaly detection:** Few studies in this literature test more than one analytical paradigm against the same dataset and the same outcome variable, largely because doing so requires carrying three separate modeling pipelines through to completion rather than reporting whichever approach proved successful first. Undertaking this comparison here revealed that the three approaches do not merely differ in performance; they disagree as to whether a usable signal ever exists. The label-blind methods returned results indistinguishable from chance, whereas the supervised method produced a result that would be considered a reasonable, if modest, predictive model within a small-sample setting. This disagreement would not have been apparent had the thesis reported only the best-performing method, and it is precisely the type of finding that a single-method study is structurally unable to produce.
- **Evidence that rejection is not necessarily equivalent to anomaly:** The Isolation Forest results, obtained across three attempts and a leakage-safe sensitivity check, indicate that Rejected shipments in this dataset are not reliably distinguishable as multivariate outliers, despite being reliably distinguishable through supervised methods that use the label directly. This constitutes a specific and testable finding rather than a general claim: it suggests that, at least for this form of shipment-level quality rejection, the underlying phenomenon does not take the form anomaly detection is designed to identify, namely a routine defect crossing a threshold, rather than a shipment that is unusual across many dimensions simultaneously. Put differently, a shipment may fail on one narrow, well-defined criterion, excess acidity, for example, without appearing multivariate anomalous in any other respect, which is precisely the profile an anomaly detector trained to flag broad multivariate deviation is least equipped to identify.
- **Evidence that data integration constraints are themselves substantively important, rather than a preliminary technical step:** Chapter 3 documented country and year contamination within the primary internal source, a supplier-

matching exercise that succeeded for only a quarter of shipments, and additional internal quality data from three further quality-control providers that could not be merged at the parameter level without losing most of its informational value. These do not constitute incidental obstacles to be resolved before the substantive analysis begins; rather, they provide a concrete, worked example of precisely the access, comparability, and validity problems that Santos et al. [1], Duan [3], and Papa et al. [9] describe as persistent, under-discussed barriers within this field. Documenting how these barriers actually manifest within one organization’s data, down to the specific match rates and merge failures reported in Chapter 3, constitutes a contribution in itself, given that the literature in which these authors are writing more often asserts that such barriers exist than demonstrates precisely what they look like within a working dataset.

6.3 Practical and Analytical Reflections

This section brings together two complementary strands of the discussion: the practical implications of the findings for Altromercato, and the analytical rationale behind the progression of approaches used to reach them.

6.3.1 Practical Implications for Altromercato

This research generates five practical implications for Altromercato.

- **Departure delay warrants monitoring as an early risk indicator:** It is the most robust multivariate predictor of rejection identified in this thesis, remaining significant alongside nine other variables in the same model, despite originating from logistics tracking, the least structured of the three axes examined here. A variable already recorded, albeit inconsistently, in the company’s shipment tracker thus carries greater standalone predictive weight than several internal quality measurements subject to far more rigorous laboratory scrutiny. This result supports the idea that consistent departure-date logging should take precedence over additional funding for quality assessment.
- **Quality variables should not be collapsed into a single score without validation:** The entropy-weighted composite score examined in Chapter 4 failed to produce a reliable representation of rejection risk, ranking Approved shipments as riskier than Rejected ones on average. Where a single risk figure is required, for instance in a supplier scorecard, it should be validated against observed outcomes

rather than assumed effective on the basis of individually relevant inputs; this thesis demonstrates that relevant inputs do not necessarily combine into a relevant score.

- **Historical rejection outcomes constitute a valuable analytical asset:** The strongest result in this thesis derives from the sole method that used the Approved/Rejected label directly. As Altromercato already records this outcome for Source of Quality 1 shipments, extending this practice consistently across all quality-control providers would be worthwhile. The inability to incorporate Source of Quality 2, 3, and 4 records into the modeling dataset reflects a data-structure limitation rather than any deficiency in the value of their outcome history.
- **Standardizing data across quality sources would substantially improve future modeling:** The failure to merge Source of Quality 2, 3, and 4 into the modeling dataset was a practical, not conceptual, limitation: providers measure different parameters, in different formats, against different thresholds, precluding direct comparison even of conceptually similar readings. Greater standardization of parameters and units across providers, where operationally feasible, would substantially improve comparability and enable a larger share of Altromercato’s existing quality-control investment, currently 83 documents excluded from the modeling dataset, to inform risk assessment rather than remain unused context.
- **Country-level context complements, but cannot substitute for, shipment-level information:** External risk indicators shifted anomaly-detection performance in a consistent direction when combined with internal data, yet remain coarse by construction, assigning a single score to every shipment from a given country regardless of supplier, region, or year. Such indicators are best treated as background context supplementing shipment-level data rather than as a substitute for it, since two shipments from the same country and year receive identical external-risk values in this framework irrespective of their actual performance.

6.3.2 Rationale Behind the Progression of Analytical Approaches

The analytical approaches presented in Chapter 4 developed progressively, shaped by the structure of the available data and by the results obtained at each stage of the analysis.

- **Starting from the available data.** The analytical choices made throughout this chapter were shaped by the dataset’s own characteristics. Information came from multiple sources, not every variable was available for every shipment, and the data itself required extraction, transformation, and integration before modeling

could proceed. These constraints, in turn, limited the type and complexity of the approaches that could reasonably be applied.

- **Exploring a composite risk representation through entropy weighting.** With several quality-deviation indicators available, the first exploratory step asked whether they could be combined into a single risk score without relying on subjective, expert-assigned weights. The Entropy Weight Method was used for this purpose, assigning weights on the basis of how much each indicator actually varied across shipments. This approach offered an initial representation of multidimensional quality risk, but the separation it produced between Approved and Rejected outcomes was not strong enough to justify retaining it as the final modeling approach.
- **Exploring multivariate unusualness through Isolation Forest.** The analysis then moved away from relying directly on the Approved/Rejected label. Isolation Forest was applied instead to examine whether shipments with unusual combinations of characteristics could be identified from the available variables. The scenario-based design further allowed the individual contributions of internal quality, logistics, and external information to be assessed progressively.
- **Testing the observed outcome through logistic regression.** Finally, since the Approved/Rejected status was known for every shipment, logistic regression was applied as a supervised approach to determine whether the selected variables could explain the observed outcome directly. Given the limited size of the dataset and the small number of Rejected observations, a comparatively simple and interpretable model was judged more appropriate than a more complex machine-learning approach.

Taken together, these exploratory steps trace the progressive development of the analytical framework. Each approach addressed the supplier risk problem from a distinct angle: entropy weighting asked whether multiple indicators could be aggregated into a common risk representation, Isolation Forest asked whether unusual observations could be identified without direct reference to the known outcome, and logistic regression asked whether that outcome could be modeled directly from the available predictors.

6.4 Limitations

The limitations of this research are discussed in detail in Section 5.3 and are summarized only briefly here.

Sample size: With 113 shipments and only 28 rejected, this dataset provides a very limited sample for training a model. This limitation is what led to more flexible modeling approaches.

Extraction accuracy was checked manually, but not exhaustively: Because this thesis’s entire modeling dataset rests on the OCR and parameter-extraction pipelines described in Chapter 3, extraction uncertainty is a fundamental methodological limitation. The pipelines were refined against known bugs and flagged failure modes, and a manual check of extracted values against source documents was carried out. However, the source documents vary considerably in layout and structure, so no single verification procedure could be applied uniformly across them all. The manual verification therefore covered a subset of cases and not a random sample of the complete dataset.

Missingness is not random: Several cut-test fields are populated for a subset of document types that does not fall across Approved and Rejected shipments, which is why the imputed datasets gave some different results.

The additional internal data could not be always used: Source of Quality 2 3 and 4 could not be merged at the parameter level without discarding most of their rows. This left a source of internal quality information outside the final model.

External data static: Every shipment from the country gets the same single-snapshot risk score no matter the supplier, region or year.

Generalization is limited: All of this comes from one company four quality-control sources and eight sourcing countries, in a small time window. Nothing can be assumed to hold without further testing.

6.5 Future Research

Multiple significant extensions for this work can be followed, building on the constraints identified above:

- **Expanding the dataset:** Increasing the dataset size would handle one of the main constraints of that research study. If we have a bigger dataset, we can also study the possibility of testing more flexible models. Special priority should be placed on ensuring additional Rejected observations. Even a small increase in the number of labeled Rejected shipments would meaningfully tighten the coefficient estimates reported in Chapter 4.
- **Standardizing quality dimensions and improving data integration:** If we place quality measurement standards for the providers, we will be able to include Sources of Quality 2, 3, and 4 in the modeling dataset at the parameter level. This

change would let us use the additional internal documents that are currently not matched as modeling observations, instead of keeping them as extra information. Also, if we assign a standard supplier and shipment ID when entering data, instead of matching by country or year, we can avoid the complex matching process described in Chapter 3.

- **Financial and Economic dimensions:** Future data collection should also include economic and financial information not sufficiently covered by the current dataset, such as procurement prices, expenses linked to rejected shipments, the costs of replacement or switching, contractual penalties, and other financial repercussions of supplier disruption. If those factors will be available, the proposed framework in this research can be extended from operational risk to a more complete assessment of the total economic risk connected with supply chain.
- **Developing a hybrid decision-support system:** A hybrid system merging expert-defined rules with data-driven models, for instance, employing departure delay and acidity thresholds as an initial expert-defined screen while a supervised model refines the assessment as further labeled data accumulate, would provide Altromercato with a usable decision support during the period preceding the availability of a dataset sufficient to support a fully data-driven model.

Bibliography

- [1] Santos, Ciro Rodrigues dos; de Oliveira, Ualison Rébula; Aprigliano, Vicente.
Supplier Risk in Supply Chain Risk Management: An Updated Conceptual Framework.
Applied Sciences, 2025, 15(13), 7128.
<https://doi.org/10.3390/app15137128>
- [2] Zsidisin, George A.; Ellram, Lisa M.; Carter, Joseph R.; Cavinato, Joseph L.
An Analysis of Supply Risk Assessment Techniques.
International Journal of Physical Distribution & Logistics Management, 2004, Vol. 34, No. 5, pp. 397–413.
DOI: <https://doi.org/10.1108/09600030410545445>
- [3] Duan, Xingyu.
Global Supply Chain Risk Management: Strategies and Mitigation Approaches in the Age of Uncertainty.
Advances in Economics, Management and Political Sciences, 2025.
DOI: <https://doi.org/10.54254/2754-1169/2025.GL28673>
- [4] Teuscher, Peter; Grüniger, Beat; Ferdinand, Niels.
Risk Management in Sustainable Supply Chain Management (SSCM): Lessons Learnt from the Case of GMO-Free Soybeans.
Corporate Social Responsibility and Environmental Management, 2006, Vol. 13, No. 1, pp. 1–10.
DOI: 10.1002/csr.81
- [5] Ivanov, Dmitry; Dolgui, Alexandre; Blackhurst, Jennifer V.; Choi, Tsan-Ming.
Toward Supply Chain Viability Theory: From Lessons Learned Through COVID-19 Pandemic to Viable Ecosystems.
International Journal of Production Research, 2023, Vol. 61, No. 8, pp. 2402–2415.
DOI: <https://doi.org/10.1080/00207543.2023.2177049>

- [6] Chapman, Pete; Clinton, Julian; Kerber, Randy; Khabaza, Thomas; Reinartz, Thomas; Shearer, Colin; Wirth, Rüdiger.
CRISP-DM 1.0: Step-by-Step Data Mining Guide.
SPSS Inc., 2000. [CRISP-DM 1.0 PDF](#)
- [7] Liu, Fei Tony; Ting, Kai Ming; Zhou, Zhi-Hua.
Isolation Forest.
Proceedings of the Eighth IEEE International Conference on Data Mining, 2008, pp. 413–422.
DOI: <https://doi.org/10.1109/ICDM.2008.17>
- [8] Lang, Sarah Maria; Zanker, Christoph; Glöser-Chahoud, Simon.
From Risk to Resilience: A Multi-Layered Framework for Supplier Risk Assessment to Strengthen Supply Chain Resilience.
International Journal of Physical Distribution & Logistics Management, 2026, Vol. 56, No. 5, pp. 574–594.
DOI: <https://doi.org/10.1108/IJPDLM-02-2025-0066>
- [9] Papa, Francesca; Grassi, Andrea; Popolo, Valentina; Vespoli, Silvestro.
AI-Driven Detection of Supply Chain Misreporting Using Engineered Cross-Stage Features and Isolation Forest.
Frontiers in Artificial Intelligence and Applications (FAIA), 2025.
DOI: <https://doi.org/10.3233/FAIA250560>
- [10] Tang, Christopher S.
Perspectives in Supply Chain Risk Management.
International Journal of Production Economics, 2006, Vol. 103, No. 2, pp. 451–488.
DOI: <https://doi.org/10.1016/j.ijpe.2005.12.006>
- [11] Christopher, Martin; Peck, Helen.
Building the Resilient Supply Chain.
International Journal of Logistics Management, 2004, Vol. 15, No. 2, pp. 1–14.
[scispace](#)
- [12] Benjamini, Yoav; Hochberg, Yosef.
Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.
Journal of the Royal Statistical Society: Series B (Methodological), 1995, Vol. 57, No. 1, pp. 289–300.
DOI:10.2307/2346101

- [13] O'Brien, Robert M.
A Caution Regarding Rules of Thumb for Variance Inflation Factors.
Quality & Quantity, 2007, Vol. 41, pp. 673–690.
DOI: <https://doi.org/10.1007/s11135-006-9018-6>
- [14] Mavi, Reza Kiani; Goh, Mark; Mavi, Neda Kiani.
Supplier Selection with Shannon Entropy and Fuzzy TOPSIS in the Context of Supply Chain Risk Management.
Procedia - Social and Behavioral Sciences, 2016, Vol. 235, pp. 216–225.
DOI: <https://doi.org/10.1016/j.sbspro.2016.11.017>
- [15] Hosmer, David W.; Lemeshow, Stanley; Sturdivant, Rodney X.
Applied Logistic Regression.
Third Edition. Wiley Series in Probability and Statistics, John Wiley & Sons, 2013.
[pdf](#)
- [16] Altromercato.
Altromercato.
Official WebSite: <https://www.altromercato.it/>

Annex A

Supplementary Data Extraction and Analysis Material

The appendices below supplement the extraction, preparation, and modeling procedures described in Chapters 3 and 4.

Source of Quality 1 Data Extraction Details

Section 3.1.1 describes the extraction pipeline for Source of Quality 1 shipment quality reports. This appendix lists the main variables that pipeline recovers from the semi-structured PDF reports, grouped by type.

Because the reports do not always follow the same uniform layout always, each parameter was located by its label rather than by a fixed position on the page: the text between one recognized label and the next was read as that parameter's value, minimum, and/or maximum, depending on how many numbers appeared in that span. Table A.1 summarizes the resulting variable groups.

Variable group	Examples	Extraction approach
Shipment metadata	Container identifier, date, country, / quality code	Located via characteristic labels and document-specific patterns
Sample information	Sample size, Shell content, moisture ,Physical characteristics	Read from the corresponding section of the document
Chemical characteristics	pH, acidity	Parameter-based extraction from the relevant text span
Defect indicators	Mould, infestation, insect damage, germination counts	Parameter-based extraction, interpreted according to the reported structure (min / max / measured)
Outcome information	Approved, Rejected	extracted from the document checked box

Table A.1: Main information extracted from Source of Quality 1 reports.

Complete Univariate Statistical Screening Results

Table 3.5 in Chapter 3 already lists the eleven candidate variables that cleared the $p < 0.05$ screening threshold. For completeness, Table A.2 below covers the remaining candidates, the ones that did not clear that threshold.

Table A.2: Not significant Variables

Variable	Information dimension	p -value	FDR q
Moisture (Umidita fave)	Internal	0.059812	0.151831
Cocoa butter (Burro cacao)	Internal	0.214784	0.393771
Germinated beans	Internal	0.229418	0.398462
Insects present	Internal	0.234512	0.404456
Double beans	Internal	0.280693	0.421039
Documentation delay	Logistics	0.809495	0.834791
Transit time	Logistics	0.145212	0.319466
Recalculated transit time	Logistics	0.145212	0.342285
Origin median transit time	Logistics	0.051767	0.142359
Discrimination risk	External	0.147620	0.304467
Civil liberties risk	External	0.195678	0.379845
Child rights risk	External	0.288715	0.414243
External risk score (overall)	External	0.288715	0.396983
Standard of living risk	External	0.390474	0.515426
Gender equity risk	External	0.390474	0.495602
Workers' rights risk	External	0.451195	0.531765
Forced labour risk	External	0.646830	0.711513
White beans	Internal	0.390958	0.477837
Shrivelled beans	Internal	0.581894	0.662156

Correlation and Multicollinearity: Numeric Tables

Section 3.4.5 and Section 3.4.6 in Chapter 3 present this screening through a heatmap figure and a short VIF summary in prose. This appendix gives the exact numbers behind both.

Variance inflation factors, final feature set

Variable	VIF
Acidity	2.493614
Infested beans	2.224822
Damaged beans	2.148146
Departure delay	1.183070
Climate & forest risk	1.185457
Partially violet beans	1.301467
Slaty beans	1.546090
Violet beans	1.359352
Mouldy beans	1.584337
Sample size	1.259873

Table A.3: Variance inflation factors

Isolation Forest: Final Refinement Scenario Results

Here are the results of the final refinement test where only the internal data extracted from Source of Quality 1 is used in the model: Table A.4 reports those scenario-level results.

Scenario	Information included	ROC-AUC
1	Source of Quality 1, significant variables only	0.253
2	Restricted internal (no additional Source of Quality 2/3/4)	0.344
3	Restricted internal + logistics	0.378
4	Restricted internal + logistics (combined)	0.48
5	Restricted internal + external	0.49
6	All dimensions combined, restricted internal set	0.52

Table A.4: Isolation Forest results

Feature-selection stability across repeated cross-validation

The repeated cross-validation also provided information on feature-selection stability. Departure delay, acidity, climate-and-forest risk, damaged beans, and infested beans were

selected in more than 80% of the folds, while the remaining variables showed lower selection frequencies.

Table A.5: Feature-selection frequency across repeated cross-validation

Variable	Selection frequency
Departure delay	50/50 (100%)
Acidity	48/50 (96%)
Climate & forest risk	47/50 (94%)
Damaged beans	46/50 (92%)
Infested beans	41/50 (82%)
Partially violet beans	33/50 (66%)
Violet beans	26/50 (52%)
Slaty beans	20/50 (40%)
Sample size	17/50 (34%)
Mouldy beans	15/50 (30%)

Annex B

Variable Definition and List of Abbreviations

Variable Definition

Variable	Dimension	Description
Approved / Rejected	Outcome	Binary shipment-level decision recorded by the company; the dependent variable throughout Chapter 4.
Acidity	Internal	Measured acidity level from the Source of Quality 1 quality report.
Damaged beans	Internal	Reported proportion or count of damaged beans.
Infested beans	Internal	Reported proportion or count of infested beans.
Partially violet beans	Internal	Reported proportion or count of partially violet beans.
Slaty beans	Internal	Reported proportion or count of slaty beans.
Violet beans	Internal	Reported proportion or count of violet beans.
Mouldy beans	Internal	Reported proportion or count of mouldy beans.
Sample size	Internal	Recorded sample size for the quality test.
Departure delay	Logistics	Actual departure date minus agreed departure date, in days.
Climate & forest risk	External	Country-level normalized risk indicator derived from the Fairtrade Risk Map.

Table B.1: Principal variables used in the final modeling dataset.

List of Abbreviations

Abbreviation	Meaning
SSCM	Sustainable Supply Chain Management
AUC	Area Under the Curve
CI	Confidence Interval
CRISP-DM	Cross-Industry Standard Process for Data Mining
DPI	Dots Per Inch
FDR	False Discovery Rate
GAC	GAC moisture measurement method
OCR	Optical Character Recognition
PDF	Portable Document Format
PNG	Portable Network Graphics
ROC	Receiver Operating Characteristic
VIF	Variance Inflation Factor

Table 2: List of Abbreviations