Building a Global Database of Supply Chain Disruptions and Their Economic Impacts

Technical Milestone Report

Benjamin Fenocchi
University of Cambridge | Supervisor: Jonathan Cullen | 18/01/2026
github.com/Ben-Fenocchi/IIB-Project

Summary.

- **Motivation:** Much of the criticality literature relies on loosely defined concepts such as supply risk and vulnerability, and on proxy indicators whose relationship to realised supply-chain disruptions has rarely been empirically tested. The absence of a structured historical record of disruption events has led to assessment frameworks that are conceptually well-structured but empirically empty.
- **Objective:** The objective of this project is to develop a quantitative, data-driven framework for assessing disruption risk in critical mineral supply chains by constructing an empirical record of physical and socio-political disruption events and using it to support validation of commonly used indicators and following that to estimate disruption likelihoods using live indicator values.
- **Approach:** The project follows a three-stage pipeline. First, a historical disruption database is constructed by filtering global news data from GDELT and extracting structured disruption events using a combination of machine-learning classifiers and large language models. Second, extracted events are validated and linked to physical and socio-political indicators, enabling empirical testing of which indicators are genuinely associated with realised disruptions. Finally, the validated indicators are used to develop predictive models for estimating live disruption likelihoods within supply chains.
- **Progress to date:** A modular processing pipeline has been implemented and exercised end-to-end on live data, enabling automated data ingestion, event extraction, and early validation. A full 20+ page, industry-standard technical documentation has also been completed and is available on the github as part of the project (*https://github.com/Ben-Fenocchi/IIB-Project*).
- **Next steps and feasibility:** Remaining work focuses on completing missing indicator retrieval, validating extracted disruption events, and developing machine-learning models to estimate disruption likelihoods from the live indicator values. Current progress demonstrates the technical feasibility of the approach and provides a clear pathway to completion.

Introduction and motivation

Critical minerals are essential to modern industrial and energy systems, yet their supply chains are often geographically concentrated and vulnerable to disruption. Demand for many such materials is projected to increase by more than an order of magnitude by mid-century, according to International Energy Agency net-zero transition scenarios.

In the past decade, critical mineral supply chains have experienced a growing number of physical disruptions arising from extreme weather events, a trend documented in supply-chain resilience surveys by the Business Continuity Institute. These can halt production at key mines, processing facilities, or transport corridors, leading to immediate supply shortfalls and downstream impacts. Such disruptions are therefore not hypothetical risks, but observed events with measurable consequences.

Despite this, there is no comprehensive structured record of realised physical disruption events affecting critical mineral supply chains. Relevant information exists across news reporting, technical bulletins, and company disclosures, but remains fragmented and unstandardised. As a result, many existing risk and criticality assessments rely on proxy indicators whose relationship to realised disruptions has rarely been empirically tested.

This project addresses this gap by constructing a structured empirical database of physical disruption events. By enabling systematic analysis of historical disruptions alongside associated indicators, the project aims to support more robust and evidence-based assessment of supply-chain disruption risk.

Objectives

The project is structured around two objectives. First, a structured historical database of realised physical disruption events is built, providing a consistent empirical record of disruption timing, location, and type. Second, predictive models are developed using indicator data to estimate disruption likelihoods, with indicator relevance and robustness assessed implicitly through model performance against observed disruption outcomes.

Scope and constraints

The project scope is restricted to natural disruptions, with an initial focus on flood-related events. While the underlying framework and database schema are designed to be extensible to other hazard types, floods are used as a primary case study due to their relevance to mining and transport infrastructure and the availability of high-quality environmental data. Disruption events are identified primarily through global news reporting, with validation supported by independent disaster databases including DFO, EM-DAT, GDACS, NASA EONET, and ReliefWeb. Flood-related environmental indicators are derived from datasets such as CHIRPS, GPM IMERG, ERA5 Reanalysis, and GloFAS, and are used to characterise both disruption and non-disruption periods.

The analysis adopts a global geographic scope and is temporally constrained to approximately the past 25 years. This window reflects the increasing availability of digital news reporting and reduced systematic reporting bias relative to earlier periods, while remaining bounded by practical processing and computational limits.

Several practical constraints apply. Continued availability of external APIs and data services is assumed, as the pipeline relies on automated retrieval from multiple providers, and dataset-specific usage restrictions must be respected, including licensing constraints on certain conflict-related sources. Computational limits are also significant: processing a single day of GDELT data requires approximately 30 minutes to filter candidate articles and several additional hours for large language model extraction, with an associated cost of roughly $1 per day of data. These constraints limit feasible temporal coverage and place practical bounds on model complexity and experimentation.

Methodology overview

The methodology centres on constructing a shared historical disruption database, developed jointly with another student, which provides the empirical foundation for subsequent modelling (division of responsibilities shown in Figure 1). This shared database defines a common schema and storage layer for disruption events derived from global news reporting and externally validated against reference datasets. All components described and analysed in the remainder of this report correspond exclusively to my individual contribution to the project.

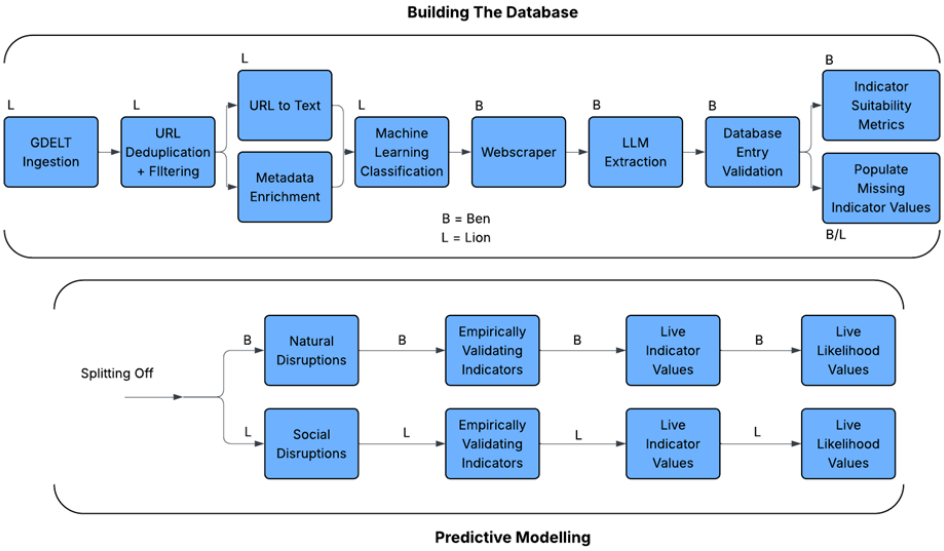


Figure 1: Pipeline overview.

Figure 1 summarises the full pipeline and delineates responsibilities. My contribution comprises the complete natural-disruption pathway and has been developed independently from scratch. This includes implementation of the web scraping stage, LLM-based disruption extraction, database validation procedures, computation of indicator suitability metrics, and population of flood-related indicator values. These components form a self-contained pipeline that supports downstream predictive modelling of disruption likelihoods.

Web scraping (implemented)

Following URL-level filtering, the web scraper converts raw article URLs into clean, normalised text suitable for downstream event extraction. Its primary role is to reliably extract article content from heterogeneous news websites while minimising systematic failures.

Given substantial variation in website structure and rendering, the scraper adopts a defensive multi-method approach. Article HTML is first processed using `Trafilatura`, which serves as the primary extraction tool and is optimised for large-scale news scraping with aggressive removal of non-content elements. For sources not reliably handled by Trafilatura, a fallback extraction using `Newspaper3k` is applied, improving coverage without introducing site-specific logic.

Extracted content is lightly normalised to remove formatting artefacts and ensure consistency. The scraper outputs a minimal structured record containing the source URL, cleaned article title, and article body text, decoupling content acquisition from downstream processing and enabling reuse without re-fetching original pages.

LLM-based disruption extraction (implemented)

Following scraping and text normalisation, articles are converted into structured disruption records using an LLM-based extraction stage. An initial hybrid rule-based and LLM approach was evaluated but proved brittle and prone to false positives from metaphorical language; the final design therefore adopts an LLM-only approach using a schema-constrained prompt.

After upstream filtering, the extraction stage processes on the order of 10^3 articles per day. Approximately 500 articles can be processed for under $1 in API costs, with runtime

and cost scaling approximately linearly with article volume. Typical runs involve $\mathcal{O}(10^5)$ output tokens, motivating strong emphasis on upstream relevance filtering to control cost and latency.

Each extracted record includes disruption type, event date and location, duration where available, extracted indicator values, supporting evidence snippets, and a model-derived confidence score. Across a representative batch of events, well-defined physical disruptions such as floods, cyclones, and earthquakes exhibit high internal consistency, with confidence scores typically in the range 0.8–0.95, while ambiguous cases are conservatively assigned lower confidence or marked as `unknown`. Confidence scores are retained as a reliability signal for downstream validation and modelling.

Database entry validation (partial implementation)

To assess the accuracy and coverage of extracted disruption events, the pipeline includes a validation stage that compares LLM-derived records against trusted external reference datasets. This component is under active development, with smoke tests implemented for all reference sources to verify data access, schema stability, and basic retrieval.

Initial validation focuses on floods and protests, using reference datasets including DFO, EM-DAT, GDACS, NASA EONET, ReliefWeb, ACLED, and MMAD. These datasets provide complementary coverage of physical hazards and social unrest and serve as external ground truth for evaluating extracted events. Additional specialist datasets will be integrated as further disruption types are introduced.

Validation is performed in two complementary directions. In forward validation, extracted events are matched against reference datasets to identify corresponding real-world events, while inverse validation checks reference events against the extracted database to assess coverage and missed events. Together, these provide insight into both extraction precision and recall.

Matching uses flexible criteria rather than exact agreement, with event dates and locations normalised or inferred where possible. Candidate matches are evaluated using approximate temporal, spatial, and textual agreement, allowing meaningful validation despite imprecise or fragmented reporting.

The validation system is implemented as a modular pipeline that standardises extracted and reference events into a common schema, generates candidate matches, and scores them using similarity measures. Outputs include matched and unmatched event lists and summary diagnostics highlighting systematic coverage gaps, which inform subsequent modelling.

Indicator suitability metrics (designed)

Following construction of the disruption event dataset, indicator suitability must be assessed prior to modelling disruption likelihood. Candidate indicators vary substantially in spatial and temporal coverage, update frequency, historical depth, and relevance to specific disruption mechanisms, and treating all indicators as equally informative risks introducing noise and bias into downstream models.

The Indicator Suitability Metrics (ISM) stage provides a systematic and transparent framework for evaluating indicator quality. Each indicator is scored against 8 metrics capturing relevance, data coverage, temporal consistency and more, enabling indicator selection to be explicit and reproducible rather than ad hoc.

ISM also supports scalability. As additional disruption types and data sources are incorporated, structured indicator scoring ensures consistent extension of the pipeline

without redesigning the modelling approach. The output is a ranked view of indicators by disruption type, restricting downstream modelling to cases with sufficient empirical support.

Population of missing indicator values (in progress)

While news-derived disruption events provide high-frequency information on when and where disruptions occur, they rarely contain complete quantitative indicator data. Many relevant environmental indicators are not explicitly reported in news articles, yet are required to construct continuous time series for likelihood modelling.

To address this, suitable external indicator datasets have been identified and assessed for populating missing values during both disruption and non-disruption periods. These datasets provide gridded environmental variables capturing background conditions before, during, and after observed events, enabling consistent comparison across time and space. Integration of these sources into the pipeline is ongoing and is a necessary step for moving from event detection toward predictive modelling of disruption likelihood, rather than reactive analysis of observed events alone.

Predictive modelling and empirical validation (planned)

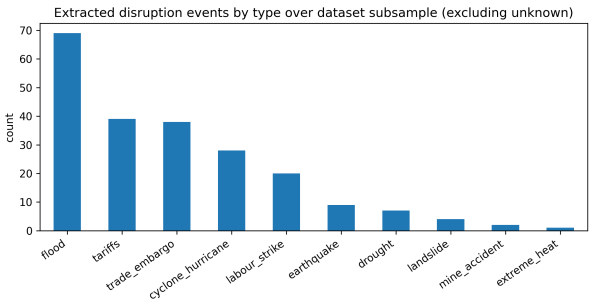
Predictive modelling would be used to assess whether the selected indicators jointly explain and predict disruption occurrence. Each observation would be represented as a feature vector constructed from indicator values at a given location and time, optionally including lagged indicators to capture lead–lag effects. Disruption and non-disruption samples would be used to form a supervised learning dataset, enabling estimation of disruption likelihoods from observed indicator behaviour.

Empirical validation would be performed implicitly through model performance. Indicators carrying predictive signal would improve discrimination between disruption and non-disruption periods, while weak or redundant indicators would be downweighted or suppressed through regularisation and out-of-sample evaluation.

Two modelling approaches are considered. A kernel Support Vector Machine would be used to capture nonlinear relationships between indicators while remaining robust in high-dimensional feature spaces, with calibrated class probabilities interpreted as disruption likelihoods. Alternatively, regularised logistic regression would provide a simpler and more interpretable baseline, with coefficient shrinkage offering a transparent measure of indicator relevance.

Achievement in the project so far

Figure 2: Extracted disruption events by type (excluding `unknown`).



The pipeline has been exercised end-to-end on daily batches of global news data. In a representative single-day run, the system processed filtered articles and produced on the order of several hundred structured disruption records spanning a range of physical and socio-political event types, including floods, cyclones, earthquakes, droughts, labour strikes, tariffs, and trade embargoes. Extracted events include normalised disruption types, inferred dates and locations, optional duration estimates, and model-derived confidence scores, demonstrating that the extraction, structuring, and storage stages are operational. Ambiguous cases are explicitly

captured as `unknown`, providing a controlled mechanism for handling uncertainty rather than discarding noisy inputs.

Limitations, risks, and known issues

The core data acquisition and extraction pipeline has been implemented end-to-end, but several components remain incomplete. Full database validation is pending, as meaningful assessment requires execution over extended historical periods, and population of indicator values for extracted events is ongoing, with suitable external datasets having been identified. As a result, predictive modelling remains future work.

A key technical risk concerns indicator availability and quality. Some indicator datasets have incomplete spatial or temporal coverage, particularly for historical periods or less well-monitored regions; this will be assessed using coverage diagnostics, with interpolation and proxy indicators used where appropriate to mitigate these effects. Indicators exhibiting weak predictive signal will be identified through ablation and sensitivity analysis, with poor performance treated as an informative outcome rather than a pipeline failure.

Further risks arise from dependence on external data sources and APIs. Rate limits, outages, or upstream schema changes may disrupt data collection, mitigated through batching, caching, and retry logic, while extraction errors or model drift may introduce noise into event records. Recall-focused design choices also increase processing cost and API exposure, motivating progressively tighter filtering as empirical performance becomes clearer.

Table 1: Remaining deliverables, execution plan, and verification criteria.

Deliverable	Purpose	Weeks	Verification
Multi-year disruption dataset	Enable meaningful validation and downstream analysis	0–2	Gap-free daily event records over multi-year historical windows
Indicator value population	Complete event-level feature coverage using external datasets	2–4	Indicator values populated for events and non-disruption periods
Validation/robustness checks	Assess extraction quality and temporal consistency	4–6	Precision/coverage diagnostics against reference sets
Predictive modelling	Establish baseline disruption likelihood estimates	6–8	Out-of-sample discrimination above random baseline
Final analysis + write-up	Consolidate results and limitations	8 +	Consistent results and reproducible pipeline

Evaluation strategy

Evaluation of the system would be performed at both the data and model levels. The database construction and extraction pipeline would be evaluated using forward and backward validation, as described in the validation methodology, to assess temporal consistency, extraction reliability, and robustness across historical periods.

The predictive model would be evaluated using backtesting, in which historical indicator values are used to generate disruption likelihood estimates for past periods and compared against observed disruption events. Performance would be assessed using out-of-sample temporal splits to ensure that results reflect genuine predictive capability rather than retrospective fitting.