Multinational Companies and Conflict: RA Notes for Data Work*

David Andrés-Cerezo
UAB and BSE

Michele Rosenberg
University of Essex

Shunsuke Tsuda
University of Essex

December 23, 2025

1 Overview

This project examines the relationship between multinational companies and conflict using multiple data sources. This note provides a high-level overview of the RA tasks; more detailed instructions will be shared in January 2026 after the RAs have been recruited.

1.1 Team Structure

The data work is primarily led by Michele and Shunsuke (PIs). There will be three RAs:

1. Two RAs will work on the S&P data (Section 2) and other data (Section 4). Reina will collaborate with the RA team based at PSE and facilitate communication with the PIs. We anticipate up to 100 hours per month in total from these two RAs from January to May 2025, though the exact workload is flexible and will be adjusted as the project progresses.

2. The third RA will focus on the GDELT database (Section 3). The PIs will first seek a candidate with prior experience working with GDELT, potentially through Upwork. If this is not feasible, the alternative will be to recruit an additional student RA from PSE. We anticipate up to 100 hours per month from January to May 2025, though the exact workload is flexible and will be adjusted as the project progresses.

The hourly wage for all RA work is £18.09.

1.2 Prerequisites & Skills

Before starting work, please review the following documents:

*E-mail: david.andres.cerezo@uab.cat, michele.rosenberg@essex.ac.uk, shunsuke_tsuda@alumni.brown.edu.

- Gentzkow & Shapiro (2014), "Code and Data for the Social Sciences: A Practitioner's Guide" ("GS2014"): read in full.
- Tsuda (2023), "Software Engineering for Social Scientists" ("ST2023"): focus on the "Portability" and "Clarity & Maintainability" sections.

Required skills:

- Proficiency in Stata for data cleaning (for S&P and other data).
- Basic Python skills (optional but helpful).
- Familiarity with QGIS and/or R for managing spatial data.
- SQL for working with the GDELT database.

Key attributes: Clarity and precision in coding, careful error minimization, and well-organized data structures.

2 The S&P Data

The raw data are sourced from S&P Capital IQ Pro. Reina cleaned the raw data following the procedure described in this note. We therefore begin our work using the resulting processed dataset. Potential tasks include the following.

2.1 Constructing company-level data

We compile all available company-level information from the S&P data, including the locations of mining assets, the timing of mine entry and exit, mining license issuance and renewal dates, and basic firm characteristics such as headquarters information. This data construction was included in the earlier procedure. Given the importance of accurate company-level information, Reina and Shunsuke will closely coordinate on this task.

2.2 Entries of MNCs (Extensive Margin)

At the mining property–year level, construct indicators capturing the entry and exit of MNCs. These include binary variables indicating whether an MNC enters or exits a given mining property in a particular year, as well as measures of years since entry and years since exit for each MNC. We also record whether a mining property experiences its first MNC entry, multiple entries, or the complete exit of all MNCs, allowing us to distinguish between initial market penetration and subsequent turnover.

2.3 Measures of Market Competition (Intensive Margin)

At the mining property–year level, construct multiple measures capturing the intensity of market competition. These include the total number of active MNCs, the number of new entrants in a given year, and concentration measures based on ownership and equity shares (e.g., Herfindahl–Hirschman indices). We also construct indicators for changes in ownership structure, such as shifts in majority control, joint ventures, and consolidation events, to capture dynamic changes in competitive pressure over time.

3 The GDELT Database

The **event-level dataset** should contain:

- Actors (all actors involved in each event)
- Actions (code and text)
- Locations (latitude, longitude, country)

Actors:

- **Companies:** All companies identified in the company-level S&P data (Section 2.1).
- **Host governments:** At all levels, including national central governments and local governments.
- **Rebel groups:** All groups identified in the ACLED data (Section 4).
- **Individuals:** For example, local civilians, foreign nationals, and businesspeople (especially in cases of one-sided violence, threats, and kidnapping or abduction). Whenever possible, record the company affiliation of businesspeople.

Actions:

- Collect all types of actions, encompassing not only violence and armed conflict but also a wide range of other activities. These include, for example, negotiations or contract agreements between multinational corporations and governments, profit-sharing disputes, threats against businesspeople by rebel groups, kidnappings or abductions, protests targeting firms, and local or national policy debates.
- First, follow GDELT's coding scheme to classify action types; more detailed classifications tailored to this project can be discussed later.
- Clearly record action types in the output data using both codes and textual descriptions.

- Whenever available, collect all quantitative information, such as numbers of casualties or deaths, numbers of fighters, ransom amounts, and the value of financial transfers.

There are many possible ways to construct a dataset from GDELT, and numerous decisions will arise along the way. Whenever a decision requires input, please communicate with us. Our goal is to collect the most detailed data possible, subject to practical constraints and trade-offs.

4 Other Data

Tasks involving other data include:

- Creating grid cells covering all of Africa to merge different types of spatial information.
- Cleaning conflict data from ACLED (Armed Conflict Location & Event Data).
- Cleaning survey data (e.g., Afrobarometer, World Bank LSMS, DHS).
- Cleaning and harmonizing other basic geographic data.
- Constructing grid cell–level variables using these cleaned datasets.