S&P Data: RA Note*

Jun Goto
GRIPS

Michele Rosenberg
University of Essex

Shunsuke Tsuda
University of Essex

April 2025

1 Project Summary

Jun Goto studies the political economy of mining and media with other collaborators. In a separate project, Michele Rosenberg and Shunsuke Tsuda examine the relationship between multinational companies and conflict using the same data, contributing to obtaining and managing the datasets.

2 Prerequisites & Skills

Before starting, please review the following documents:

- Gentzkow & Shapiro (2014) "Code and Data for the Social Sciences: A Practitioner's Guide" ("GS2014" henceforth): Read fully.
- Tsuda (2023) "Software Engineering for Social Scientists" ("ST2023" henceforth): Focus on the "Portability" and "Clarity & Maintainability" sections.

Required skills:

- Proficiency in STATA for data cleaning.
- Basic Python skills (optional but helpful).
- Familiarity with QGIS and/or R for managing spatial data.

Key attributes: Strive for clarity and precision in coding, minimize errors, and maintain well-organized data structures.

*E-mail: gotojun.jg@gmail.com, michele.rosenberg@essex.ac.uk, shunsuke_tsuda@alumni.brown.edu.

3 Code & Data

3.1 Code

Store all code in code/clean. **Code should be as self-documenting as possible.** Follow the guidelines in the "Clarity and Maintainability" section of ST2023 and the provided sample codes (code/clean/sample_...xxx.do).

Update the **master code** (code/master.do) accordingly. The master code serves two primary purposes:

1. To execute all scripts in the correct sequence.
2. To document the structure and sequence of research-related code, detailing the purpose of each script and any additional required information, such as STATA packages needed to run the entire program.

As such, please avoid creating separate documentation files (e.g., Word or text files) to explain your code. Instead, include all necessary documentation within the master code itself. Detailed guidance is in the "Reproducibility" section of ST2023 and Chapter 2 of GS2014.

Inputs/Outputs:

- Input raw data from data_S&P/matals_mining/...
- Export cleaned data to data/raw_cleaned/S&P_cleaned
- Temporary files can be stored in data/temp

3.2 Raw Data (Input)

The primary source of raw data is S&P Capital IQ Pro, which provides detailed market intelligence across industries. We downloaded this data in June 2024, and it is stored in our shared Dropbox folder: Project_S&P/data/raw/data_S&P.[1]

We focus on metals and mining because this sector offers substantially richer data relevant to our research objectives. Data for this sector is located in the subfolder matals_mining.

There are numerous Excel files with extensive variable documentation, outlined in the following files:

- data_S&P/S&P_data_variables_metals_mining_properties.docx contains information on all the data in the following subfolders:
  - data_S&P/matals_mining/properties_property_details
  - data_S&P/matals_mining/properties_production

[1] The folder contains PDFs describing the data platform. You are not required to read them.

  - data_S&P/matals_mining/properties_mine_econ_modeled_data
  - data_S&P/matals_mining/properties_reserves_resources
  - data_S&P/matals_mining/properties_technical_geology
  - data_S&P/matals_mining/properties_financings
  - data_S&P/matals_mining/properties_most_recent_transactions
  - data_S&P/matals_mining/properties_top_drill_results

- data_S&P/S&P_data_variables_metals_mining_claims.docx contains information on all the data in the following subfolder:
  - data_S&P/matals_mining/claims

- data_S&P/S&P_data_variables_metals_mining_drill_results_capital_costs.docx contains information on all the data in the following subfolder:
  - data_S&P/matals_mining/drill_results_capital_costs

- data_S&P/S&P_data_variables_metals_mining_transactions.docx contains information on all the data in the following subfolder:
  - data_S&P/matals_mining/transactions

3.3 Data to be Constructed (Output)

We need six types of datasets:

1. Property level (cross-sectional data with **time-invariant information**)

   Key: property ID [prop_id]

2. Property-year level (panel data with **time-variant information**)

   Keys: property ID, year [prop_id, year]

   * Please generate the year ID (year) by yourself. The raw data contain multi-period information in the wide format. It is necessary to reshape the data to the long format.

3. Company-level (cross-sectional data with time-invariant information)

   Key: company ID [company_id]

   * Please generate the company ID (company_id) by yourself. In the raw data, different data files use different variables to represent the company ID, including:
     - operator_snl_instn_key in property_details_3_operator_1_{region}.xls
     - owner_snl_instn_key in property_details_5_ownership_details_1_{region}.xls

- historical_owner_snl_instn_key
  in property_details_6_historical_ownership_details_1_{year}_{region}.xls

  There might be other data files that contain the company ID using a different variable name. Please check and generate the same company ID in such cases as well.

4. Company-year level (panel data with time-variant information)

   Keys: company ID, year [company_id, year]

5. Grid cell level (cross-sectional data with time-invariant information)

   Key: grid cell ID [cell_id]

6. Grid cell-year level (panel data with time-variant information)

   Keys: grid cell ID, year [cell_id, year]

Requests:

- Use STATA for 1.-4. For 5.-6., discuss further before proceeding.
- Include **all** time-invariant information in the cross-sectional datasets and **all** time-variant information in the panel datasets.
- Ensure clear distinctions between datasets and avoid redundant variables:
  - Avoid including any time-invariant variables (other than the key) in any panel dataset.
  - Please do not include the same variable (other than the key) across different datasets.

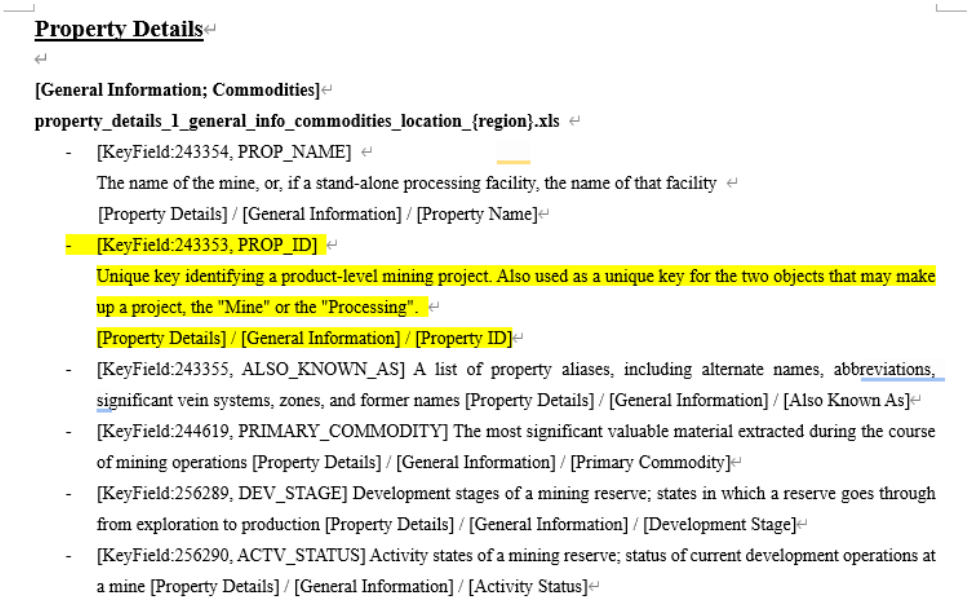
  Refer to Chapter 5 of GS2014 and understand the role of keys in organizing data.
- Clearly define variable names and provide concise labels (80-character limit in STATA). Generate labels by importing words from the corresponding documentation, focusing on the most critical details:
  - For example, Figure 1 (from data_S&P/matals_mining/properties_property_details) highlights some variables from the dataset, written in bold.
  - For the variable marked yellow, the variable name will be `prop_id` and the label will be such as "`Unique key identifying a product-level mining project.`"
  - Variable names can be in lowercase; STATA may automatically convert them when importing Excel files.
- Please ensure that each output file name clearly specifies whether it contains cross-sectional or panel data.

- **1. & 2. (the property-level datasets)** Please export datasets based on the subfolders classified in section 3.2:
  - `properties_property_details_crosssection.dta`
    `properties_property_details_panel.dta`
    `properties_production_crosssection.dta`
    `properties_production_panel.dta`
    ⋮
  - `claims.dta`: Data form TBD
  - `drill_results.dta`: Data form TBD
  - `transactions.dta`: Data form TBD
- **3. & 4. (the company-level datasets)** Similarly, please export the following datasets:
  - `companies_crosssection.dta`
    `companies_panel.dta`

* The desired details for the output data structure can be discussed further as we progress with the work.

Tips:

- For raw Excel files spanning multiple regions, append all regions into a single dataset for ease of processing.



Property Details

[General Information; Commodities]

property_details_1_general_info_commodities_location_{region}.xls

- [KeyField:243354, PROP_NAME]
  The name of the mine, or, if a stand-alone processing facility, the name of that facility
  [Property Details] / [General Information] / [Property Name]
- [KeyField:243353, PROP_ID]
  Unique key identifying a product-level mining project. Also used as a unique key for the two objects that may make up a project, the "Mine" or the "Processing".
  [Property Details] / [General Information] / [Property ID]
- [KeyField:243355, ALSO_KNOWN_AS] A list of property aliases, including alternate names, abbreviations, significant vein systems, zones, and former names [Property Details] / [General Information] / [Also Known As]
- [KeyField:244619, PRIMARY_COMMODITY] The most significant valuable material extracted during the course of mining operations [Property Details] / [General Information] / [Primary Commodity]
- [KeyField:256289, DEV_STAGE] Development stages of a mining reserve; states in which a reserve goes through from exploration to production [Property Details] / [General Information] / [Development Stage]
- [KeyField:256290, ACTV_STATUS] Activity states of a mining reserve; status of current development operations at a mine [Property Details] / [General Information] / [Activity Status]

Figure 1: Example