Data Report. Title: Battle-related deaths during natural resource conflicts in Latin-America from 1946 to 2006

1. Questions

- What is the procentage of civillians being killed during natural resource conflicts?
- Top 5 years in which the largest percentage/number of civillians in Latin-America during natural resource conflicts were killed?

---

2. Data Sources

2.1 Descriptions of Data Sources

Datasource Name	Description	Availability	Data Type	Geographic Coverage
The Natural Resource Conflict Dataset	Codes whether internal armed conflicts are clearly linked to natural resources from 1946 - 2006.	[1]	Stata	Global, covering multiple regions
UCDP Battle-Related Deaths Dataset version 24.1	Contains data on battle deaths (soldiers and civilians killed in combat) in state-based conflicts for 1946–2008.	[2]	Zipped CSV	Global, covering multiple regions

2.2 Structure and Quality of Data Sources

Table 1. First row of The Natural Resource Conflict Dataset

acdid	conflepid	epstartdate	ependdate	begin	end	location	ccode	sidea	sideb	re
1	1_1946	1946-06-01	1946-07-21	1946	1946	Bolivia	145.0	Bolivia	Popular Revolutionary Movement	0.

1. **The Natural Resource Conflict Dataset** in a Table 1

Step of Data Cleaning	Description
Unnecessary Columns	Removed columns: 'acdid', 'sidea', 'sideb', 'ccode', 'res_confl', 'aggrav', 'finance', 'distribution', 'epstartdate', 'ependdate'.
Remove Empty Rows	Used the pandas method dropna.
Remove Duplicates	Used the pandas method drop_duplicates.
Correct Errors	There were no misspellings in the data.
Standardize Formats	Transformed 'start_date' and 'end_date' into new columns: 'whole_days', 'days', 'months', 'years' for simpler calculations. Dropped original date columns as they became unnecessary.

Step of Data Cleaning	Description
Improve Data Quality	Kept only rows for Latin-American countries: 'Bolivia', 'Paraguay', 'Costa Rica', 'Guatemala', 'Cuba', 'Argentina', 'Venezuela', 'Colombia', 'Dominican Republic', 'Peru', 'El Salvador', etc.
Manual Filtering	Performed manual filtering for Latin-American countries due to the absence of a region column, unlike datasets such as "UCDP Battle-Related Deaths Dataset version 24.1".

Table 2. First row of UCDP Battle-Related Deaths Dataset version 24.1

conflict_id	dyad_id	location_inc	side_a	side_a_id	side_a_2nd	side_b	side_b_id	s
13324	14236	Kyrgyzstan, Tajikistan	Government of Kyrgyzstan	132		Government of Tajikistan	131	

1. **UCDP Battle-Related Deaths Dataset version 24.1** in a Table 2

Step of Data Cleaning	Description
Unnecessary Columns	Removed columns: 'conflict_id', 'dyad_id', 'side_a_id', 'side_a_2nd', 'side_b_id', 'side_b_2nd', 'territory_name', 'type_of_conflict', 'battle_location', 'gwno_a', 'gwno_a_2nd', 'gwno_b', 'gwno_b_2nd', 'gwno_loc', 'gwno_battle', 'version'.
Remove Empty Rows	Used the pandas method dropna.
Remove Duplicates	Used the pandas method drop_duplicates.
Correct Errors	There were no misspellings in the data.
Standardize Formats	This dataset only included a 'year' column with a numeric value.
Improve Data Quality	Kept only rows for Latin-American countries by filtering using a region column where value 5 indicates the Americas region, which was retained after filtering.
Automatic Filtering	Filtered the dataset automatically by using the region column, where value 5 corresponds to the Americas region, and dropped the column after the filtering operation.

2.3 Licenses and Permissions

Since both datasets, The Natural Resource Conflict Dataset and UCDP Battle-Related Deaths Dataset version 24.1, belong to The Peace Research Institute Oslo (PRIO), we could look at licences and permission denoted by this organization. At this web page you could find this information:

`...open data policies benefit Norwegian social science – and PRIO in particular...All data collection efforts at PRIO are openly available...` from [3]. This proves that they are under a standard open-data license. Additionally, it was included that, in order to use theirs datasets for a research purpose, you need to include the references to these papers, where they were first presented. You could cite it from [1] and [2].

2.4 Data Pipeline

ETL (Extract, Transform, Load) Process:
The data pipeline was implemented for two datasets using Python. It consists of several steps provided below:

Pipeline Name	Extract	Transform	Load
Natural Resource Conflict Dataset Pipeline	Reads metadata and downloads Stata files.	Filters rows, removes unnecessary columns, standardizes data format, filters manually for Latin-America countries, removes empty cells and duplicates, adds a `years` column from `start_date` and `end_date`.	Stores transformed data in SQLite database as `conflicts.sqlite`.
UCDP Battle-Related Deaths Dataset Pipeline version 24.1	Reads zipped metadata and downloads CSV files.	Filters rows, removes unnecessary columns, filters automatically for Latin-America countries, renames `location_inc` to `location` for consistency with the first pipeline, removes empty cells and duplicates.	Stores transformed data in SQLite database as `deaths.sqlite`.

More information regarding which transformations and cleaning steps were done and why you could find in a previous section `2.2 Structure and Quality of Data Sources`.

I encountered with the problems:

- when I tried to filter my datasets for Latin-America countries without a region information, which made me to search for the region of the country manually using maps;
- when I tried to standardize data format into 'years', 'days', 'months' and 'whole days' from data string type;
- to decide which columns are meaningful for my task.

Regarding Handling Errors or Changing Input Data, I used:

- `prevent_errors` function to remove all empty rows and duplicates using pandas methods: `dropna` and `drop_duplicates`
- I changed input data of the dataset 1, which is a datatype string, in order to achieve consistency with the dataset 2. Finally we got a new column `year` to check correspondence with each database.

---

3. Results and Limitations

- **Results**:
  - Datasets stored in SQLite include only Latin-America countries. They were stored in a that way due to necessity of combining 2 different tables, that could be done using foreign keys, e.g. `year` and `location`.
  - The datasets are complete by removing empty and duplicates values
  - They are consistent in terms of same columns of a `year` and a `location` across both datasets

    - They have same timeliness in terms of the `year` column.
- **Limitations**:
    - Timeliness issue: The dataset 2 [2] does not include the starting and ending dates, but rather `years`. For this reason, we could depend regarding precise days on the data given from the table 1 [1].

References

1. The Natural Resource Conflict Dataset- Web page.
2. UCDP Battle-Related Deaths Dataset version 24.1 - Web page.
3. PRIO license.
4. UCDP Battle-Related Deaths Dataset version 24.1 - Link to the article.