Data Report by Yuliia Herasymenko

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

1. **Datasource1: The Natural Resource Conflict Dataset**
    - The Natural Resource Conflict Dataset code whether internal armed conflicts are clearly linked to natural resources from 1946 - 2006.
    - It is available here [1].
    - The data being used through this project could be found here [2].
    - Data Type of the dataset is a Stata, which covers global data, e.g. multiple geographic regions.
2. **Datasource2: UCDP Battle-Related Deaths Dataset version 24.1**
    - A dataset on battle deaths (number of soldiers and civilians killed in combat) in state-based armed conflicts for the period 1946–2008.
    - It is available here [3].
    - The data being used through this project could be found here [4].
    - Data Type of the dataset is a zipped CSV, which covers global data, e.g. multiple geographic regions.

2.2 Structure and Quality of Data Sources

Table 1. First 2 rows 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	1_1967	1967-03-01	1967-10-16	1967	1967	Bolivia	145.0	Bolivia	ELN	0.

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

- Data Cleaning:
  - The unnecessary columns were removed, s.a. 'acdid', 'sidea', 'sideb', 'ccode', 'res_confl', 'aggrav', 'finance', 'distribution', 'epstartdate', 'ependdate'.
  - Remove empty rows: pandas method `dropna`
  - Remove duplicates: pandas method `drop_duplicates`
  - Correct errors: There were no misspellings
  - Standardize data formats: 'start_date' and 'end_date' would be transformed into columns with names 'whole_days', 'days', 'months', 'years' for simplicity of calculations. Later on these original columns would be dropped due to nonnecessity
  - Improving the data quality: The table was cleaned by including only Latin-America countries, such as 'Bolivia','Paraguay','Costa Rica','Guatemala','Cuba','Argentina','Venezuela','Colombia','Dominican Republic','Peru','El Salvador','Uruguay','Surinam','Panama','Trinidad and Tobago','Haiti','Mexico','Nicaragua'. Unfortunately, I filtered it manually, because the dataset does not provide a region column which includes countries from particular geographic region, as it was with "UCDP Battle-Related Deaths Dataset version 24.1".
    - Data Imputation:
    - No missing values being found

Table 2. 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	
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
   - Data Cleaning:
     - The unnecessary columns were removed, s.a. '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: pandas method `dropna`
     - Remove duplicates: pandas method `drop_duplicates`
     - Correct errors: There were no misspellings
     - Standardize data formats: this dataset only includes column 'year' with a numeric value
     - Improving the data quality: The table was cleaned by including only Latin-America countries. Hopefully, I filtered it automatically, because the dataset provides a region column which includes countries from particular geographic

region, where value 5 includes Americas region, which would be deleted after a filtering operation
    - Data Imputation:
    - No missing values being found

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:

`In the same way as international organizations and international order benefit a small country like Norway, 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 non-commercial or research purpose, you need to include the references to these papers, where they were first presented.

For the UCDP Battle-Related Deaths Dataset version 24.1 use the link to the project description website [6]. However, you could also find this information of how to refer to these datasets by downloading an appendix or a cookbook via [1] and [3].

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:

1. **Pipeline's structure for a Natural Resource Conflict Dataset**

- **Modules**:
    - **Extract**: Reads metadata and downloads Stata files.
    - **Transform**: Filters rows, removes unnecessary columns, standardize data format, filter manually Latin-America countries, remove empty cells and duplicates, add `years` column from `start_date` and `end_date`.
    - **Load**: Stores transformed data in SQLite database as conflicts.sqlite.

1. **Pipeline's structure for UCDP Battle-Related Deaths Dataset version 24.1**
    - **Modules**:
    - **Extract**: Reads zipped metadata and downloads CSV files.
    - **Transform**: Filters rows, removes unnecessary columns, filter automatically Latin-America countries, rename `location_inc` to `location` for a consistency with the first pipeline, remove empty cells and duplicates.
    - **Load**: 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.

ETL Diagram:

Data Sources (From Metadata) --> Extract --> Transform --> Load --> Database (SQLite)

---

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 [3] 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. The Natural Resource Conflict Dataset - Data link.
3. UCDP Battle-Related Deaths Dataset version 24.1 - Web page.
4. UCDP Battle-Related Deaths Dataset version 24.1 - Data link.

5. PRIO license.
6. UCDP Battle-Related Deaths Dataset version 24.1 - Link to the article.