Assessing the impacts of extreme weather events on the Socioeconomics – A case study in Southeast Asia region

1 Introduction

1.1 Motivation for selecting Southeast Asia region

Nowadays, Southeast Asia has emerged as a region of significant development and economic success. However, Southeast Asia faces a daunting challenge: it is one of the world's most vulnerable to the impacts of climate change and natural disasters[1]. This vulnerability comes from different factors, including low-lying land susceptible to rising sea levels, high frequency of floods and droughts, large populations at risk, heavy reliance on agriculture for economic stability and limited community resilience to withstand climate shocks. Emerging Southeast Asian countries as the Philippines, Thailand, Indonesia, Vietnam have experienced some of the most extreme weather events globally. Accordingly, this data project aims to provide insights into how extreme weather events have affected the socioeconomics of this region.

1.2 Questions that interest the audience

To investigate the impact of extreme weather events on the socioeconomic landscape of Southeast Asia, this project focuses on two key questions:

"What are the patterns in the occurrence of extreme weather events and the socioeconomic status of Southeast Asian countries?". To explore the question, the report will later reveal critical insights, including the typical types of extreme weather events since 1902, the countries with the highest number of fatalities, and the countries experiencing the highest frequency of extreme weather events.

"How do extreme weather events impact the socioeconomic landscape of Southeast Asian countries?" Based on the World Risk Index, we will examine the relationship between disaster risk indicators and various socioeconomic factors, including income, prevalence of undernourishment, life expectancy, unemployment rate, sanitation, and injuries.

2 Datasets

Two datasets were initially identified to fulfill the project research: Southeast Asian (SEA) extreme weather events and global socioeconomic indicators. Upon closer inspection, these datasets however were deemed insufficient for subsequent analysis due to lack of meaningful indicators in extreme weather events dataset and there are lots of missing values. Accordingly, disaster risk dataset from the World Risk Index (WRI) was incorporated to augment the research. This brought the total numbers of datasets employed in this project to three. The datasets are provided in CSV format to facilitate seamless integration and analysis within Python ETL workflows. Python was chosen over Jayvee due to its extensive library support for accessing the Kaggle API and handling CSV files.

2.1 Southeast Asian extreme weather events

The dataset is available to the audience through the following source:

- Metadata URL: https://data.opendevelopmentmekong.net/dataset/disaster-in-southeast-asia-from-1900-to-2021
- Data URL: https://data.vietnam.opendevelopmentmekong.net/dataset/a6232bd8-c77e-40f7-ab1b-5fe824de52ce/resource/d1b3b83a-4312-44e7-806d-8f16bf83ed3a/download/disaster_south_eastern_asia_en.csv
- Data Type: CSV
- License: CC BY-SA 3.0 DEED License

This dataset, derived from The International Disaster Database (EM-DAT), was refined to include a total of 2,737 extreme weather events in Southeast Asia from 1902 to 2021. It contains originally 52 standard attributes. During

[1] According to UN Women "Southeast Asia belongs to one of the most disaster-prone regions in the world". https://wrd.unwomen.org/explore/regions/southeast-asia-asean

data transformation tasks, the most 26 relevant attributes were selected for this project. Descriptions of these attributes are provided in the repository at /project/project-plan.md.[2]

2.2 Southeast Asian socioeconomic indicators

The dataset aggregated from World Bank Open Data source is currently available on Kaggle:

- Metadata URL: https://www.kaggle.com/datasets/mjshri23/life-expectancy-and-socio-economic-world-bank
- Data URL: https://www.kaggle.com/datasets/mjshri23/life-expectancy-and-socio-economic-world-bank
- Data Type: CSV
- License: World Bank Dataset Terms of Use

This dataset includes 19 years data of multiple countries with the following attributes: Country, Country Code, Region, Income group, Year (2001-2019), Life expectancy, Prevalence of undernourishment, Carbon dioxide emissions, Health expenditure, Education expenditure, Unemployment, Corruption, Sanitation, Disability-Adjusted Life Years (DALYs) due to Injuries, Disability-Adjusted Life Years (DALYs) due to Communicable diseases, Disability-Adjusted Life Years (DALYs) due to Non-Communicable diseases. The dataset provides significant socioeconomic indicators and contains minimal missing values, playing a crucial role in addressing the given questions.

For the specific use case in this project, the dataset was narrowed down to include only the Southeast Asian region.

2.3 Southeast Asian disaster risk

The dataset was aggregated from World Risk Index (WRI) Meta Data as follows:

- Metadata URL: https://data.humdata.org/dataset/worldriskindex?
- Data URL: https://data.humdata.org/dataset/1efb6ee7-051a-440f-a2cf-e652feccf73/resource/3a2320fa-41b4-4dda-a847-3f397d865378/download/worldriskindex-trend.csv
- Data Type: CSV
- License: Creative Commons Attribution International

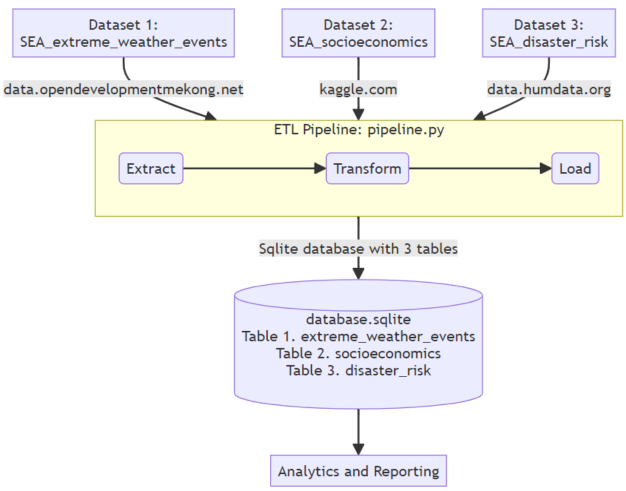
The WRI dataset integrates a country's physical exposure to extreme weather events with its societal vulnerability. Exposure analysis encompasses earthquakes, cyclones, floods, rainfall, droughts, and climate-induced sea-level rise. Societal vulnerability is assessed based on susceptibility to extreme weather events, lack of coping capacities, and limited adaptive capacities. All components of the index are normalized on a scale of 0 to 100, with higher scores indicating a greater national risk of disaster.

The dataset includes 11 years data of multiple countries with the following attributes: Region, WRI, Exposure, Vulnerability, Susceptibility, Lack of Coping Capabilities, Lack of Adaptive Capabilities, Year (2011-2021), Exposure Category, Vulnerability Category, Susceptibility Category.

3 Methodology

For this project, three datasets were obtained from various online sources and subsequently integrated into a SQLite database, as illustrated in the diagram on the right.

The pipeline starts by extracting data from three different online sources. Afterward, specific transformations are applied to satisfy analytic requirements, as detailed in Section 3.1. In the final step, the transformed data is stored in SQLite database, where it can be accessed for later use.



3.1 Extract, Transform, Load (ETL) Pipeline

The following table provides a detailed description of the ETL tasks performed in this project by pipeline.py. Two

[2] Audience can also review the attribute descriptions of the EM-DAT Database on the official website: https://doc.emdat.be/docs/data-structure-and-content/emdat-public-table/

customized functions, *extract_csv_data* and *extract_data_from_kaggle*, were utilized to download data from online sources and extract it to CSV format. During the transformation task, built-in functions were applied to rename columns, refill missing values, drop unused columns, limit the dataset to countries in Southeast Asia region. These adjustments are necessary for both socioeconomics and disaster risk data, as the datasets were intended to provide global indicators. Following data imputation for the two datasets, there are no remaining gaps or missing values.

ETL Pipeline	Dataset	Task
Extract	SEA extreme weather events	Apply function `extract_csv_data` to download the csv dataset from source Link to download
	SEA socioeconomic indicators	Apply function `extract_data_from_kaggle` to download zip file and extract csv dataset from kaggle. The function can be reused for every Kaggle dataset resources Link to download
	SEA disaster risk	Reuse function `extract_csv_data` to download the csv dataset from source Link to download
Transform	SEA extreme weather events	Rename country name to ensure consistency with two other datasets, drop and change columns name, refill missing values
	SEA socioeconomic indicators	Rename columns name, filter the dataset to include only countries within the Southeast Asia region, refill missing values
	SEA disaster risk	Rename and drop columns name, filter the dataset to include only countries within the Southeast Asia region
Load	SEA extreme weather events	Apply function `to_sql` to load 3 transformed datasets into a storage destination, namely database.sqlite. In the terminal run the command `*py project/pipeline.py*`
	SEA socioeconomic indicators	
	SEA disaster risk	

Table 1. Tasks performed in ETL pipeline

3.2 Results and limitations

For this project, I utilize SQLite database due to its lightweight, and no configuration requirements. As a result of ETL pipeline, SQLite database named 'database.sqlite' has been generated, containing three tables. The overview of datasets is accessible through SQL commands at the respository /project/data-exploration.ipynb.

There are some limitations that could have impacted the later analysis:

- Temporal and spatial limitations: in extreme weather event dataset, numerous missing values (over 1,000 values) were identified in the following attributes *Aid Contribution, Local Time, Associated Dis, Appeal, Declaration, Origin, Dis Mag Value, Event Name, Geo Locations.* To maintain data accuracy, attributes with many missing values should not undergo data imputation by using mean, mode, median. In the project K-Nearest Neighbors Imputation were initially executed in the transformation task to predict local time when an extreme weather event occurred. This results in a predominant occurrence of 11:10 (hours, minutes) local time in the attribute, which is however considered unreliable. Predicting event local times could introduce significant errors in subsequent analyses as the information is often unique, random, and context-specific for each event. This observation also applies to Event Name, Geo Locations and other attributes.
- Inconsistency in Year attribute: in extreme weather event the year ranges from 1902 to 2021, while in the disaster risk it ranges from 2000 to 2023. Integrating these datasets with others may present temporal challenges due to this discrepancy and the lack of comprehensive information.