Data Report:
Impact of Extreme Weather Events on the USA

vorgelegt von:

Richard Heinz

Course of Study: M.Sc. Informatik

Course: MADE

Project Work: 4

24. November 2024

Inhaltsverzeichnis

1 Introduction **1**

2 Data Sources **1**
2.1 NOAA Storm Events Database 1
2.2 Bureau of Economic Analysis: GDP by State 2

3 Data Pipeline **3**

4 Result and Limitations **3**

1 Introduction

This study develops an interest in the investigation of the economic impacts of extreme weather events on the United States as a result of increasing storms, floods and tornadoes. The overriding question, therefore, remains:

> **What economic impact do weather extremes like storms, floods and tornadoes have on the USA?**

In this study, analysis of historical weather events is done in light of economic indicators to understand the extent of damage events create to regional economies and identify areas most prone to severe economic loss. The analysis hereby provides a broad knowledge and facts useful for decision-makers in ensuring better preparedness against disasters or economic resilience.

2 Data Sources

2.1 NOAA Storm Events Database

The core data source for the analysis will be from the NOAA Storm Events Database, which contains detailed records of extreme weather events across the United States. Every entry found in the database carries with it comprehensive details about the kind of weather event, location, duration and the impact of the event. Attributes included in the structure will also demonstrate not just the timing but the severity of each weather event. Following is a sample of the data reduced to the most relevant columns from a CSV-File for the year 1999:

Tabelle 1: Overview of Weather Events

Event	State	Begin	End	Dmg	Impact
Ice Storm	South Carolina	1999-01-02	1999-01-03	$1.54M	A shallow cold arctic airmass caused a significant ice storm in Upstate South Carolina. Freezing rain and sleet led to downed trees, power lines, and widespread outages affecting 160,000 people.
Hail	Illinois	1999-01-21	1999-01-21		Hailstorm caused minor damage in Herrin, Illinois, with localized impact.

Firstly, it should be noted that the column names have been modified for improved readability. Regarding the quality of the NOAA Storm Events Database, it is evaluated based on the following criteria:

- **Accuracy**: Most of the data are considered reliable, obtained from weather reports, authorities, and trained spotters. However, there are some discrepancies in damage estimates or event severity because reporting may have been inconsistent, especially in rural or less-monitored areas, and underreporting or misclassification might affect accuracy.
- **Completeness**: While the database encompasses most of the events, smaller ones or those in less populated areas could be underreported or missing, especially in the *Damage (Dmg)* column. Records may also be incomplete with older records due to reporting standards in the past.
- **Consistency**: The data structure is uniform, aiding analysis. However, inconsistencies may occur in event classification and damage reporting, such as varying conventions for “Flood“ versus “Flash Flood“ complicating comparisons.

- **Timeliness**: The NOAA database includes both current and historical data. However, delays are common, particularly for large events, as in the case of Hurricane Helene, which is still being assessed and thus lacks full data processing.
- **Relevancy**: The dataset is highly relevant to this study, focusing on extreme weather events with economic impacts. However, relevancy may decrease if event types or data collection methods change, requiring periodic updates.

The NOAA Storm Events Database is issued under a standard open-data license, which calls for proper citation when the dataset is used. The dataset is freely available for electronic download. By the terms of this license, I will provide appropriate acknowledgement of the source when using the data in my work, and follow additional guidelines provided by NOAA. The full terms and conditions of use can be viewed on the website https://www.ncei.noaa.gov/metadata/geoportal/rest/metadata/item/gov.noaa.ncdc:C00510/html inside the tab *Constraints*.

2.2 Bureau of Economic Analysis: GDP by State

BEA data comes with economic indicators including GDP, Personal Income, and Employment by State. Data is divided into categories like Real Dollar Statistics and Current Dollar Statistics. The time period covered is from the year 1999 to 2023, and it includes both national and state-level details, enabling the comparative study of economic trends.

The structure is clear, with each row representing a specific economic indicator and columns capturing yearly data points. However, some of the data fields, such as Real personal incomeänd Regional price parities,"do have missing values, which likely indicates that data collection or estimation was incomplete for these variables. This therefore might point to limitations regarding the full assessment of some economic aspects.

- **Accuracy**: The BEA dataset originates from reliable governmental sources and therefore is usually quite accurate. Missing values, such as in 'Real personal income' and 'Regional price parities,' may limit analysis of some economic indicators.
- **Completeness**: The dataset provides a wide-ranging view of economic indicators at both the national and state levels. However, many fields have missing data in them, reflecting incomplete records, especially in real-dollar statistics, reducing the depth of historical comparisons.
- **Consistency**: The data structure is consistent throughout the dataset, following a uniform format for each economic indicator and across the years. This consistency facilitates straightforward analysis. However, data definitions may differ slightly for certain variables between states, requiring careful interpretation during analysis.
- **Timeliness**: The dataset provides a historical overview from 1999 up until 2023, allowing for trend analysis over time. However, the timeliness is limited for studies requiring the ongoing year 2024, as it is not finished yet and thus there are no final results.
- **Relevancy**: This dataset has a high relevance to economic analyses, particularly in understanding the dynamics of regional economies and trends related to GDP, income, and employment. The data aligns with the objectives of the study by focusing on the economic performance across various states over time.

Data from the Bureau of Economic Analysis is available to the public and is, therefore, in the public domain. Thus, it has no restrictions for use. BEA does require that users attribute their use, which constitutes their rules for the public domain. So as not to be in violation of these requirements, I will cite them appropriately throughout the analysis where their information is utilized. The Information can be accessed via the Webiste https://www.bea.gov/open-data#:~:text=BEA%20makes%20its%20statistics%20available,major%20data%20tools%20and%20datasets..

3 Data Pipeline

The data pipeline has been implemented with Python, using the key libraries *Pandas*, *BeautifulSoup*, *Selenium*, and *Requests*. This data pipeline automates downloading, extracting, and transforming data from NOAA Storm Events Database, and the Bureau of Economic Analysis data for the GDP by state.

Data Cleaning and Transformation Several transformation and cleaning steps ensured the data usability. The NOAA data, which was downloaded in compressed CSV format, was extracted, combined and stored in Excel format, with each year occupying a separate sheet. Unnecessary columns like injuries and deaths were removed. For the BEA data, the extraction involved Website-Interaction using a Selenium-driven browser, ensuring proper download of the CSV files, which were previously validated. The need Selenium resulted from the fact that the data from BEA cannot be downloaded directly via the URL, as the download link is generated by JavaScript. The columns of the dataset correspond to the years and have already been reduced via the website filter. Further reductions in relation to the rows will result from the correlation analysis in the next step.

Challenges Encountered and Solutions A number of challenges arose during implementation. For the NOAA data, it was necessary to parse the website for download links, as the data was made available in separate files based on the year of occurrence. Therefore, all necessary files had to be downloaded, unpacked and summarized for the following analysis. The BEA website offers a download function for the dataset, where the file is dynamically generated via JavaScript. To ensure reliable navigation to the download button, Selenium's explicit wait conditions were employed. This approach allowed the pipeline to handle dynamic elements effectively, ensuring that downloads were initiated only after the correct elements were fully visible.

Meta-Quality Measures and Error Handling To handle data quality issues and ensure robustness, several meta-quality measures were implemented. Any failed downloads or extraction steps are logged, allowing for a quick identification of problematic files. In case of incomplete or broken data, subdirectories are checked at the start of the pipeline, and data is re-downloaded and transformed if necessary. Furthermore, the pipeline checks for inconsistencies in column formatting, raising warnings if unexpected data patterns are detected. The use of Excel format for storing NOAA data simplifies error inspection and makes it easier to compare historical data over time. This ensures that the pipeline can handle changes in input data structure or updates to the dataset format with minimal disruption.

4 Result and Limitations

The output data of the pipeline in regard to the NOAA Storm Events consists of a cleaned and transformed Excel file in the *.xlsx* format. It is structured in several sheets by year. The accuracy is generally good, as it represents the events during a year, including their duration, damages, and a textual explanation of the event. However, some entries may be missing or incomplete. The output data of the pipeline also includes the BEA GDP dataset in CSV format. The NOAA data is organized by year, whereas the BEA data is tabular, with economic indicators across several years. Both datasets are generally accurate, but gaps in data, especially for newer years, and incomplete records may hinder the final analysis. Another limitation that could impact the analysis is the potential exclusion of smaller, unrecorded, or missing events. These events, while potentially affecting the U.S. economy, may not be reflected in the correlation between GDP and storm events, leading to a possible underestimation of their economic impact.