2024

Curation and analysis of AI ready environmental justice datasets : a proof-of-concept study

Paridhi Parajuli

CURATION AND ANALYSIS OF AI READY ENVIRONMENTAL JUSTICE DATASETS: A PROOF-OF-CONCEPT STUDY

Paridhi Parajuli

A THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science
in
The Department of Computer Science
to
The Graduate School
of
The University of Alabama in Huntsville

May 2024

Approved by:

Dr. Tathagata Mukherjee, Research Advisor
Dr. Tathagata Mukherjee, Committee Chair
Dr. Chaity Banerjee Mukherjee, Committee Member
Dr. Sundar Christopher, Committee Member
Dr. Letha Etzkorn, Department Chair
Dr. Rainer Steinwandt, College Dean
Dr. Jon Hakkila, Graduate Dean

Abstract

CURATION AND ANALYSIS OF AI READY ENVIRONMENTAL JUSTICE DATASETS: A PROOF-OF-CONCEPT STUDY

Paridhi Parajuli

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science

Computer Science

The University of Alabama in Huntsville

May 2024

Equity and Environmental Justice (EEJ) advocates for unbiased distribution of environmental impacts across communities, regardless of social and economic characteristics. After extreme events like natural disasters, EEJ gains importance due to evident disparities in impact among communities. Addressing these injustices requires comprehensive datasets and analytical methods for quantification and resolution. While AI and advanced data analysis offer promising solutions, creating AI-ready EEJ datasets is challenging due to heterogeneity in the data surrounding EEJ. In this work, we focus on curating novel datasets for EEJ targeting a few recent extreme events - Maui Wildfire, Hurricane Harvey, and Hurricane Ida. We demonstrate the utility of the datasets using preliminary analysis with machine learning and AI enabled methods. Succinctly, we created masks to identify EEJ issues and generated nuanced insights employing machine learning, image processing and statistical methods. This study has the potential to empower authorities in data-driven policy-making, disaster management, and resource allocation, addressing the actual needs of affected communities.

Acknowledgements

I extend my profound gratitude to my advisor **Dr. Tathagata Mukherjee** for his invaluable guidance and unwavering support throughout the course of this thesis. His expertise and insightful feedback have been instrumental in shaping the direction of my research and academic growth.

I am deeply appreciative of **Dr. Manil Maskey** for introducing me to this project, providing a platform for me to contribute to impactful research initiatives. I extend special thanks to Dr. Rajat Shinde for his meticulous guidance at a granular level, contributing significantly to the refinement of both the vision and execution of this thesis.

My sincere appreciation also goes to **Mr. Iksha Gurung** and **Mr. Muthukumaran Ramasubramanian** for their valuable support and insights, enhancing the depth and quality of my work. Lastly, I express my heartfelt thanks to my friends and family for their unwavering support, encouragement, and understanding throughout this academic journey. This thesis stands as a collective achievement, made possible by the collaborative efforts and mentorship of these remarkable individuals.

Table of Contents

List of Figures

List of Tables

Chapter 1. Introduction

1.1 Background

As we navigate to creating a equal and just society, the concept of environmental justice takes the center stage. The term Equity and Environmental Justice (EEJ) refers to the equal distribution of environmental burdens and benefits to all the communities irrespective of their racial, social, demographic and economic characteristics. Environmental injustice becomes more pronounced in the wake of extreme and hazardous events, whether they stem from natural occurrences or human-made causes [1]. From the disparate impacts resulting from a wildfire to the health inequalities stemming from the geographical placement of industries, and the uneven consequences brought about by specific rules and regulations, environmental justice manifests in various forms and dimensions. This is what makes it a sensitive and intricate subject, given the challenge of identifying instances of injustices and charting a course for resolution. The complexity is heightened by its direct connection to the lives of individuals, making it a matter that directly impacts communities. Therefore, it is crucial to address this issue in order to promote an equal distribution of resources to all the communities as per their need. However, the above-mentioned issues can be addressed to a significant

extent using a data-driven framework with the use of Artificial Intelligence and in-depth data analytics [2].

1.2 Equity and Environmental Justice Data Components

When addressing Equity and Environmental Justice concerns, various viewpoints can be explored, including issues emerging post-natural disasters, disparities in policy implementation impacting communities differently, challenges specific to certain occupations, and consequences linked to developmental patterns. This study specifically focuses on the perspective of natural disasters, exploring how various natural events affect communities with diverse racial, gender, and socioeconomic characteristics in terms of damage and recovery. We have compiled novel EEJ datasets focused on natural disasters that encompass all the important dimensions of EEJ. In addition to the socio-economic and demographic features of communities, the dataset also contains complementary data that directly or indirectly support the EEJ narrative. This includes information about the disaster, as well as remote sensing data depicting the pre- and post-disaster scenarios. The components of our data set are described below.

1.2.1 Events

The curated dataset focuses on three major events, known for their severity and significant impact on people's lives. To ensure diversity, we selected a wildfire event and two hurricane events, carefully choosing the Area of Interest (AOI) based on the fire and hurricane landfalls, respectively. The Maui Wildfire of

2023, Hurricane Harvey in 2017 and Hurricane Ida in 2021 were chosen as event instances for the EEJ data curation and case studies for our analysis.

- The **Maui Wildfire** ignited fully on August 8, 2023 and rapidly escalated, attributed to dry conditions, consuming over 17,000 acres and resulting in a devastating toll of over 100 lives lost and 60 severe injuries [3]. According to the Federal Emergency Management Agency (FEMA), estimated capital loss incurred due to Maui wildfire is nearly $5.5 billion along with substantial damage to over 2,200 buildings. This catastrophe profoundly affected vegetation, critical infrastructure, and economic activities, evident in a surge of unemployment claims from 130 to 2,705 cases per week [4]. Unveiling instances of environmental injustice exacerbated by the fire presents a unique challenge, necessitating a comprehensive understanding of the wildfire's far-reaching consequences.

- The **Hurricane Harvey** struck the east coast of Texas on August 25, 2017, bringing over 60 inches of precipitation to the southeast region and causing extensive flooding [5]. The storm resulted in 68 casualties, power outages affecting 336,000 customers, and an estimated total cost of $125 billion according to National Oceanic and Atmospheric Administration (NOAA). This research focuses on examining the disparities in damages brought by the hurricane in various communities in the affected area.

- The **Hurricane Ida** made landfall near Port Fourchon, Louisiana on August 29, 2021 as a category 4 hurricane resulting in total fatalities of 107

and an estimated damage of $75.3 billion [6]. To demonstrate the various perspectives of accessing EEJ, we take this event to analyse the disparities in the recovery post disaster over the time, over different zip-codes and their community characteristics.

1.2.2 Remote Sensing Data

High-resolution remote sensing data play a crucial role in addressing EEJ issues. Given that EEJ issues inherently have both spatial and temporal dimensions, remote sensing data serve as a valuable resource for conducting analyses in both dimensions. We leverage the disparities in remote sensing data captured before and after a disaster to assess the extent of damage caused. Additionally, remote sensing data serve as effective proxies for variables that may be physically challenging to measure within the short pre and post timestamps. Here are some remote sensing sources that we have used.

- **Sentinel 2**: It is a multi-sprectral data consisting of 13 bands covering various spectral ranges from visible and near-infrared to shortwave infrared. The spatial resolution varies across bands, with 60 meters for Bands 1 and 9, 10 meters for Bands 2, 3, 4, and 8, and 20 meters for Bands 5, 6, 7, 8a, 11, and 12. These bands are combined to compute different land cover indices. We leverage these indices as proxies for estimating damage caused by events such as fire or flooding. The calculated indices include the Normalized Difference Vegetation Index (NDVI), Normalized Difference Built-up

Index (NDBI), Normalized Difference Water Index (NDWI), and Normalized Difference Moisture Index (NDMI) [7].

NDVI = (B08 - B04) / (B08 + B04)

NDMI = (B08 - B11) / (B08 + B11)

NDWI= (B03 - B08) / (B03 + B08)

NDBI = (B11-B8) / (B11+B8)

Apart from sentinel -2 data products, there are sentinel 5 data products that are used to measure air quality measure including aerosols, ozone and other gases [8].

- **Visible Infrared Imaging Radiometer Suite (VIIRS)**: VIIRS imagery captures spectral bands, including visible and infrared wavelengths [9]. Specifically, we utilized VIIRS night-time lights data, of spatial resolution 375m, as a proxy for economic/industrial activity. This data helped us study how economic activities change before and after disasters. Additionally, Fire Information for Resource Management System (FIRMS) also provides access to the VIIRS active fire detections.

1.2.3 Census Data

The U.S. Census Bureau conducts a decennial census every 10 years, supplemented by the more frequent American Community Survey (ACS) conducted annually [10]. The data is organized hierarchically, spanning state, county, tracts, and census block groups. Our study leverages both census and ACS data, utilizing

a variety of variables at the tract and block levels where available. This includes demographic information such as race, gender, housing, and income, which is later integrated with event data for in-depth analysis.

1.2.4 Pre-existing Environmental Justice Data

Different private and public authorities like the US Environmental Protection Agency (EPA), Centers for Disease Control and Prevention (CDC), NASA are working in the EEJ domain. Any freely available open datasets related to environmental justice can be used in combination with the disaster, remote-sensing and census data.

1.3 Motivation

The motivation for this research initiative comes with the Justice 40 initiative by the federal government of the US, under which 40% of the overall benefits of certain Federal investments flow to disadvantaged communities that are marginalized, underserved, and overburdened by pollution [11]. Following an objective to identify communities overburdened by environmental conditions, we propose various image processing and machine learning approaches that can serve as an initial dataset in addressing this issue. We believe the curation of comprehensive datasets in the environmental justice domain will empower researchers to conduct spatial and temporal analyses on the aftermath and recovery processes following a disaster. We aim to fill the gap in the absence of a unified dataset and EJ analysis by open-sourcing AI-ready datasets representing the EJ aspects

of a disaster in a three-dimensional cube. This research serves as a proof of concept for curating EJ datasets and introduces a set of methods and techniques for conducting preliminary analyses on the data. The outcomes of this study will aid in enhancing disaster preparedness, management, and data-driven policy making. Apart from promoting equity and equality, this study will help concerned authorities in urban planning and resource allocation based on demand.

Chapter 2. Literature Review

This section comprises of the discussion of related works and research efforts made so far in addressing Environmental Injustice issues.

2.1 Previous Works

One of the earliest researches in environmental justice was by Freeman *et al.* where cross tabulation was performed to see correlations between pollution exposure with income and race across different US locations [12]. During 1980s, research in environmental justice mostly focused on advocacy of political ideas, lacking scientific quality. However, after 1990 the field advanced methodologically - leveraging complex scientific techniques [13].

The availability of comprehensive datasets that can represent most of the dimensions spanning the domain is crucial for enabling rigorous scientific research. As demonstrated by a study [14] by Rasp *et al.*, highlights the importance of an unified dataset for weather forecasting, encompassing various climate variables. Another research for a benchmark data creation is illustrated by Betancourt *et al.* focusing on tropospheric ozone levels prediction leveraging historical air quality data and weather station's metadata. This study employs innovative proxies, such as night-time light data for gauging industrial activity and population density for

assessing human emissions [15]. Inspired by these initiatives, we draw parallels to the necessity of curating comprehensive datasets for EEJ domain. While these studies [15, 14] establish a benchmark dataset for their respective domains, they do not delve into the dimensions of environmental justice. Our work seeks to address this gap through incorporating community demographic, socioeconomic information alongside temporal geospatial data.

Ho *et al.* explored the role of social-demographic, land cover, human mobility and built environment features in shaping disparities in exposure to urban heat, flood and air pollution across 6 US counties [16]. Employing interpretable machine learning algorithms like random forests and gradient boosting, they predicted hazard risk using the social-demographic, land cover, human mobility and built environment features as illustrated by Figure 2.1.The Analysis revealed that social-demographic features play a predominant role in influencing the disparities observed in hazard exposure, shedding light on environmental injustice.

A notable research effort was conducted by from Nunez *et al.* in identifying county-level racial/ethnic and socioeconomic inequalities in emissions changes from the span of 1970-2020 after the enforcement of the Clean Air Act in US. From the association analysis of change in emissions with racial and socioeconomic features shown in Figure 2.2, it revealed association in the relative decrease in emissions among different racial and socioeconomic groups, empirically highlighting environmental injustice issues.

Prior research efforts centered around analyzing inequalities across space and time where as event-centric spatio-temporal based approaches remain under

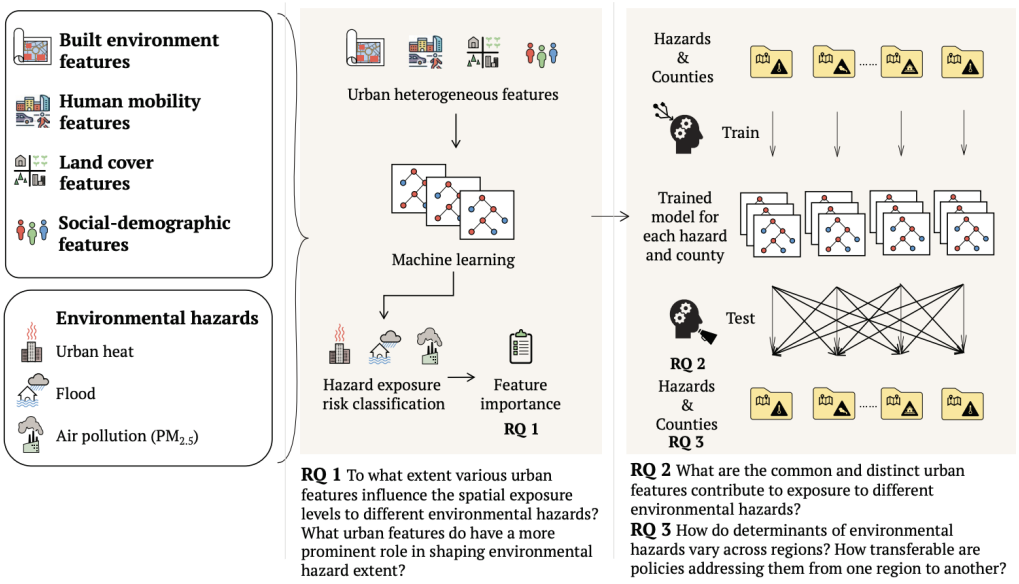

Figure 2.1: Study Overview for Relationship Urban Features and Environmental Justice using Interpretable Machine Learning by Ho *et al.*

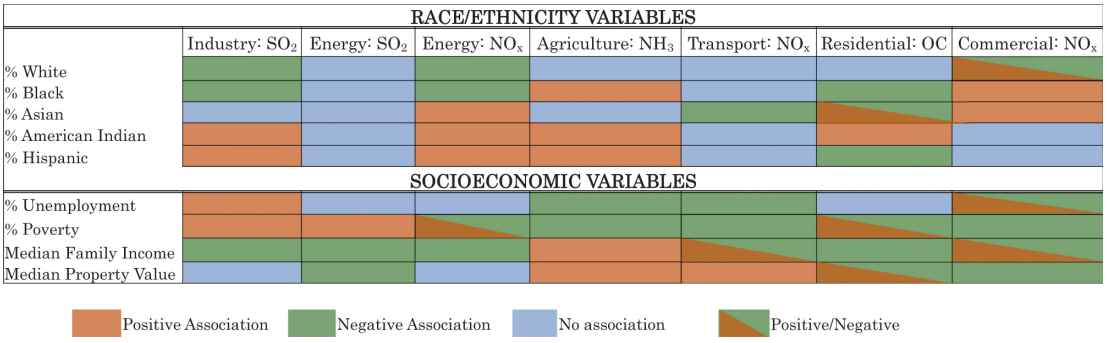

Figure 2.2: Summary of Results for Association Study by Nunez *et al.*

explored. Event based EJ analysis are crucial in accessing how different events like a global pandemic or a natural disaster affect different communities. As exemplified by a EJ study of the event COVID-19 by Segovia-Dominguez *et al.*, introduces a consensus ML model to investigate the relationship among severity of COVID 19 with air quality across communities of different socioeconomic backgrounds [17].

The effects of a natural disaster go beyond structural damages and loss of lives. It impacts different communities in diverse ways and unfortunately, it is the socially and economically vulnerable communities who feel the most significant repercussions[18]. Numerous studies regarding the aftermath of devastating disasters like hurricanes have highlighted how they have affected vulnerable population like people with disabilities [19]. This kind of research sheds light on how disproportionately disaster are affecting communities and outlines the need for effective disaster management and preparedness. Additionally, there is evident heterogeneity in the recovery process following a catastrophic event [20]. Certain areas take shorter time to recover while certain areas do not recover for prolonged amount of time. Understanding the role of demographic and socioeconomic features in creating such disparities in the recovery time would be helpful in optimizing the effectiveness of the disaster management process.

As far as calculating damages brought by disaster is concerned, there has been multiple studies that suggest remote sensing data such as change in land cover, night time lights as good estimator for the extent of damage [15, 21] . An innovative example of leveraging remotely sensed data is illustrated by the work

from NASA IMPACT team where they detected blue tarps as a damage measurement variable after Hurricane Ida happened in New Orleans [22]. Through advanced image processing techniques on high resolution Planet Lab Imagery of the area of impact, they detected blue tarp over different months following the disaster. They identified heterogeneity in the rate of change of blue tarp detections over time for different zip codes. Validation of this methodology was uniquely conducted by cross-referencing news articles to derive damage severity in each zip code, thereby comparing it with the severity obtained from blue tarp detection. Figure 2.3 shows the blue tarp detections over a specified area in New Orleans , five months post Hurricane Ida. However, a socioeconomic study of zip codes exhibiting slower versus faster recovery rates has not been conducted, presenting a valuable opportunity for an extension to this work.

2.2 Pertinent Theoretical Concepts

This section discusses the core theoretical concepts used in the research.

2.2.1 Correlation

Correlation is a widely used statistical measure and quantifies the relationship and association between two variables. Among the various correlation coefficients, Pearson's Correlation Coefficient assesses the linear relationship between variables. It is calculated using the formula:

$$\rho_{XY} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y},$$

Figure 2.3: Blue Tarp Detection Over New Orleans on February 12, 2022.

where

ρ_{XY}: Pearson's correlation coefficient,

$\text{cov}(X, Y)$: covariance with X and Y,

σ_X: variance of X,

σ_Y: variance of Y

The value of Pearson's correlation coefficient ranges from -1 to 1, where high positive values signify strong positive correlation and low negative values signify strong negative correlation. Apart from Pearson's coefficient, Spearman's rank correlation and Kendall's tau are alternative methods used for assessing the strength and direction of monotonic relationships based on the ranks of the data points [23].

2.2.2 Random Forest

Random Forest is a tree-based machine learning algorithm that creates an ensemble of multiple decision trees to perform a classification or regression task. Decision trees work by recursively splitting the data based on the features that have the most predictability for the target variable [24]. The main principle of decision trees is to choose split features that have the maximum information gain and minimum entropy regarding the predictability of the target variable. However, decision trees are very sensitive to the training data and are prone to overfitting.

Random Forest, as the name suggests, is a collection of decision trees trained on random samples of the data with a random subset of features, illustrated by Figure 2.4. This process of performing random sampling of data and features is called bootstrapping. Whenever it has to make predictions, the output of all trees is combined and presented as a prediction. For a random forest regression task, where the predicted variable is a continuous variable, it uses aggregation on the different outputs of the trees and presents it as the final output. Since tree-based methods are known to capture the non-linearity of data, random forest works well for non-linear complex data with many features [25].

The performance of a random forest regressor model can be assessed using metrics like mean square error that calculates the average squared values of the error in predictions. Similarly, explained variance ratio can give us information about the fit of the data, explaining how much variance in the target variable can be explained by the predictor variables. A random forest model can be utilized using the implementation available in the scikit-learn library.

2.2.3 K-means Clustering

K-means clustering is a unsupervised machine learning technique that aims to create clusters of similar data points based on distance/similarity between them [26]. The technique aims to minimize the intra-cluster distances and maximize the inter-cluster distances. Hence the data within the same cluster should be as homogeneous as possible, whereas the data across different clusters should be as heterogeneous as possible [24]. In practice K-means clusters are computed

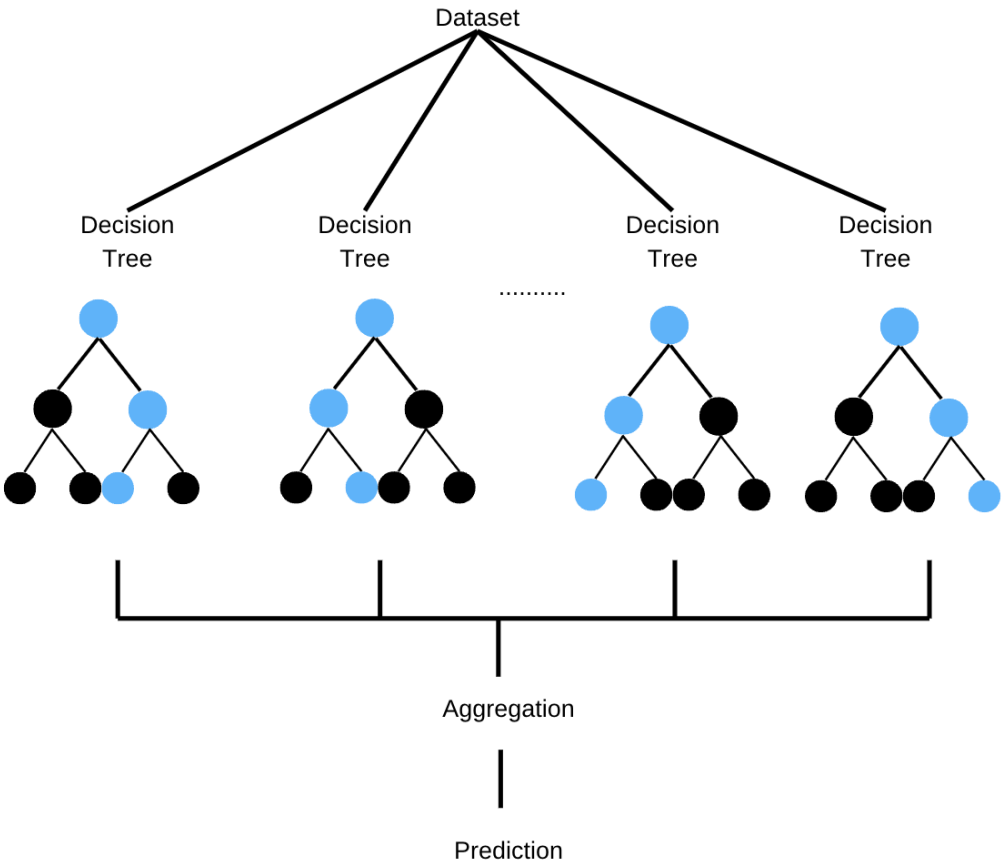

Figure 2.4: Random Forest as Collection of Decision Trees.

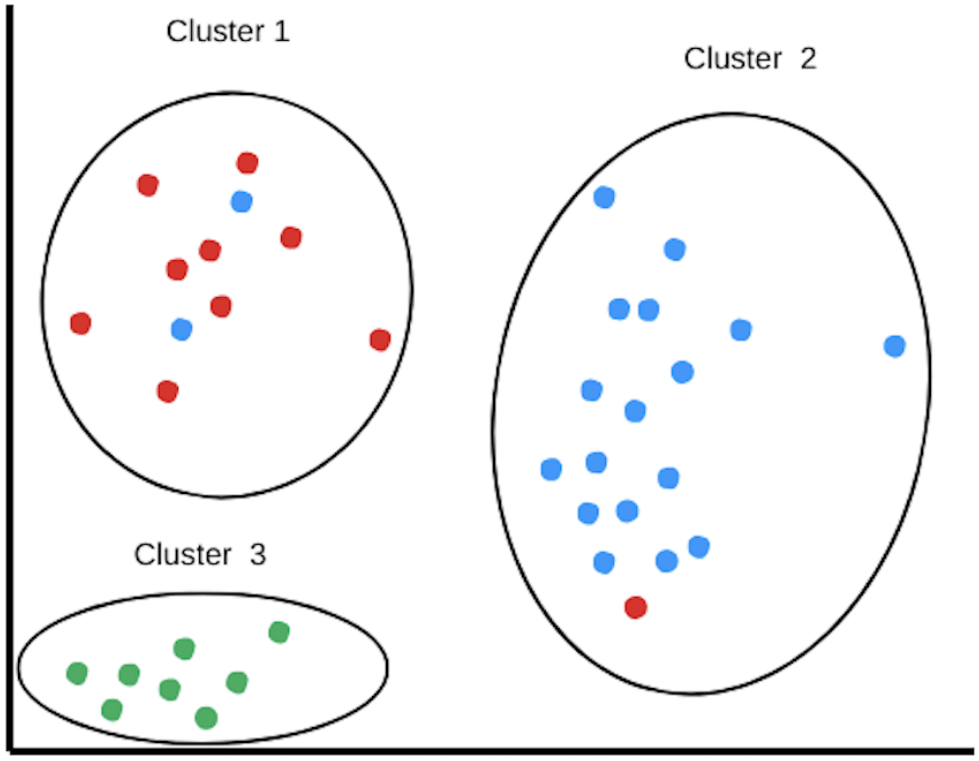

Figure 2.5: Illustration of K-means Clustering.

using an iterative algorithm that tries to homogenize the computed clusters over iterations and stops when no further progress can be made.

The performance of K-Means clustering can be evaluated using the inertia value, which represents the sum of squared distances within the clusters. Additionally, the silhouette score and davies bouldin score are also used to assess how similar an object is to its own cluster compared to other clusters. K-means clustering algorithm can also be used for feature extraction [27], image segmentation and labeling data for semi supervised learning [28].

2.2.4 Segment Anything Model

Segment Anything Model (SAM) [29] is an image segmentation model developed by Meta AI pre-trained on huge amount of data. It has the ability to segment out specified objects from an image with reasonable accuracy. It has also been found good at zero shot learning i.e.,the models ability to segment a new image that it has not been trained on [30]. This can open up a possibility of using SAM for segmenting images and labeling them so that semi supervised learning can be performed for domains that lack labeled dataset [31].

2.2.5 Data Centric AI

Data-Centric AI (DCAI) [32] emphasizes prioritizing data and its quality to drive AI outcomes, rather than focusing solely on the model. Adhering to the principle that the quality of input data significantly influences the quality of outcomes, DCAI encompasses the implementation of processes and algorithms designed to enhance results, particularly when working with smaller datasets. In conjunction with the rise of foundation models [33] in AI, this approach allows us to address challenges in domains where a unified dataset is lacking. By using foundation models initially to label data and applying DCAI principles to enhance dataset creation, significant progress in machine learning can be made for areas lacking dedicated datasets.

Chapter 3. Dataset

In this section, we delve into the details of curating comprehensive datasets for Maui Wildfire, Hurricane Harvey, and Hurricane Ida in a format that is user-friendly and conducive to analysis. Our datasets aim to encapsulate the disaster along with its associated damage and community data, providing an Equity and Environmental Justice (EEJ) perspective through the disaster-in-a-cube analogy.Our study on EEJ with respect to disaster aftermath and recovery can be summarized by the data pipeline shown in Figure 3.1.

3.1 Maui Wildfire Data Cube Creation

This section discusses the process involved in creation of AI ready data cube for Maui Wildfire.

3.1.1 Data Collection

The dataset used for the Maui Wildfire EEJ analysis was curated using remote sensing data and US census data. Different categories of data were collected for the data cube preparation is listed in Table 3.1

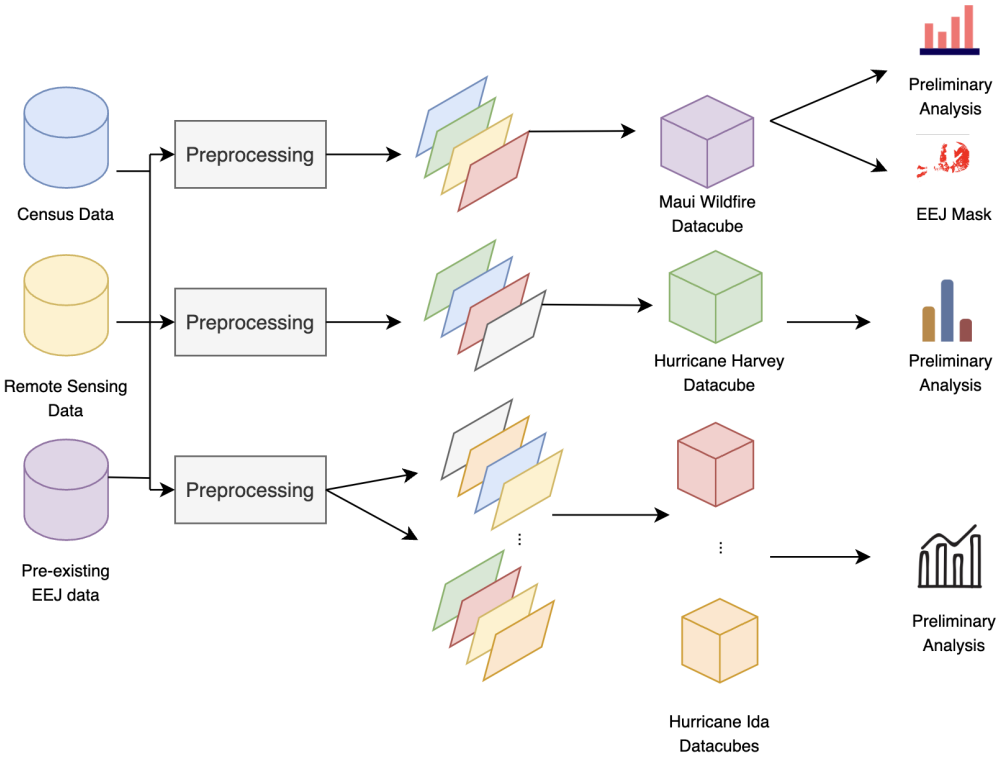

Figure 3.1: Data Processing and Curation Pipeline Overview.

Table 3.1: Description of different data sources and its attributes for Maui Wildfire.

Category	Attribute	Frequency
Racial (US Census)	Percentage of White, Black, African American, American Indian and Alaska Native, Asian, Native Hawaiian, Pacific Islander, Hispanic	Constant
Gender (US Census)	Percentage of men and women	Constant
Household (US Census)	Percentage of household 1- 7	Constant
Economic (US Census)	Median household income	Constant
Land Cover (Copernicus Sentinel-2)	NDVI, NDBI, NDMI, NDWI	Pre/Post
Socioeconomic (VIIRS)	Night time light	Pre/Post
Air Quality (Copernicus)	Aerosol Index	Pre/Post
Fire (FIRMS)	Active Fire Detections	During

3.1.2 Data Processing

The data comes from different sources, each with distinct spatial resolution and coordinate reference systems (CRS) [34]. So preprocessing is required to stack these datasets into a data cube. The general preprocessing steps applied to create the Maui Wildfire data cube were as follows.

- **Rasterization**: The data that were available as raster data were made into a uniform file format i.e., the geotiff file format [35]. For vector data like the US census data, we utilized vector to raster conversion tools and brought them to geotiff file format as well.
- **Re-projection**: Given the different sources of data, they come in different projection and follow different Coordinate Reference Systems (CRS). All of the raster data were brought to same CRS i.e., EPSG:4326 for further processing.

- **Clipping**: After identifying the Area of Interest (AOI) for the wildfire, we clipped the raster data based on the identified AOI. The AOI corresponding to the event is described below by the GeoJSON projected to EPSG:4326.


```
{
  "type": "Polygon",
  "coordinates": [
    [
      [-156.70353317911116, 20.62256935688882],
      [-156.70353317911116, 21.06701380111118],
      [-156.1815887348888, 21.06701380111118],
      [-156.1815887348888, 20.62256935688882],
      [-156.70353317911116, 20.62256935688882]
    ]
  ]
}
```


- **Resampling**: The raster data should be downsampled/upsampled to the same dimension i.e., 512*512 grid in order to be stacked into a data cube. During upsampling, nearest interpolation method was used for discrete data like fire pixels where as linear interpolation was used for continuous data like nighttime light.

3.1.3 Calculating Derived Channels

From the different raster data obtained from source data, we derived a set of raster data required for our study and hence went into our data cube.

- **Fire Proximity**: This was derived from the active fire detections raster data. It was calculated such that each pixel value in fire proximity data has the distance to its nearest fire pixel from the active fire detection. A lower value indicates close proximity to fire where as higher value indicates farther in proximity to fire.

- **Difference**: The difference between NDVI, NDMI, NDBI, NDWI, Night time light data and Aerosol Index from pre and post timestamps were calculated and added as ***Variable***_post - ***Variable***_pre. These difference data were used as proxies for damage due to the wildfire.

3.1.4 Description of Maui Wildfire Data Cube

The data cube is stored in a NetCDF [36] file format, which is a widely used file format for geo-spatial data storage, with 3 dimensions : x, y and channel representing longitude, latitude and the channel name respectively in EPSG:4326 projection. The actual data is stored as data variable name *band_data* which by selectable using the channel name.
The dims of the data cube NetCDF file is : *(x:512, y:512, channel:37)*
The channel names and description of the channels is in Table 3.2 and Figure 3.2 illustrates a subset of channels present in the data cube.

Table 3.2: Description of Different Channels in Maui Wildfire Data Cube.

Channel Name	Description
NDVI_post	Normalized Difference Vegetation Index calculated after the wildfire
NDWI_post	Normalized Difference Water Index calculated after the wildfire
NDBI_post	Normalized Difference Built-up Index calculated after the wildfire
NDMI_post	Normalized Difference Moisture Index calculated after the wildfire
NDVI_pre	Normalized Difference Vegetation Index calculated before the wildfire
NDWI_pre	Normalized Difference Water Index calculated before the wildfire
NDBI_pre	Normalized Difference Built-up Index calculated before the wildfire
NDMI_pre	Normalized Difference Moisture Index calculated before the wildfire
NDVI_diff	Difference in Normalized Difference Vegetation Index after wildfire
NDWI_diff	Difference in Normalized Difference Water Index after wildfire
NDBI_diff	Difference in Normalized Difference Built-up Index after wildfire
NDMI_diff	Difference in Normalized Difference Moisture Index after wildfire
nt_post	Night-time light detections after the wildfire
nt_pre	Night-time light detections before the wildfire
nt_diff	Difference in night-time light data after wildfire
fire_mask	Fire detections
fire_proximity	Fire proximity map
AI_pre	Aerosol Index before wildfire
AI_post	Aerosol Index after wildfire
AI_diff	Difference in Aerosol Index after wildfire
H013002 - H013008	Percentage of household size 1, 2, 3, 4, 5, 6, 7+ respectively
B19013_001E	Median Household Income
P003002 - P003008	Percentages of white, black, Indian, Asian, Hawaiian, other race, and 2 or more races
P012026	Percentage of women population
P012002	Percentage of men population

3.2 Hurricane Harvey Data Cube Creation

This section discusses the process involved in creation of AI ready data cube for Hurricane Harvey.

3.2.1 Data Collection

The dataset used for the Hurricane Harvey EEJ analysis for Hurricane Harvey was curated using remote sensing data, US census data and openly available EJ datasets like diseases and healthcare accessibility data. Different categories of data were collected for the data cube preparation is listed in Table 3.3.

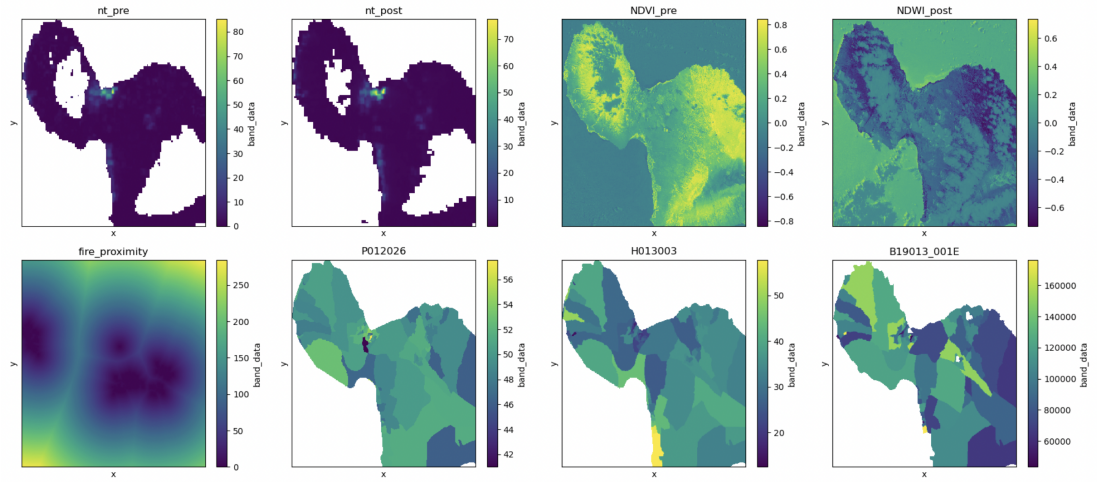

Figure 3.2: Illustration of Selected Channels from Maui Wildfire Data Cube.

Table 3.3: Description of different data sources and its attributes for Hurricane Harvey.

Category	Attribute	Frequency
Health Accessibility [37]	Accessibility to Healthcare	Constant
Disease (CDC) [38]	Percentage Population with Chronic Diseases	Constant
Topography (Copernicus)	Digital Elevation Model (DEM)	Constant
Racial (US Census)	Percentage of White, Black, African American, American Indian and Alaska Native, Asian, Native Hawaiian, Pacific Islander, Hispanic	Constant
Gender (US Census)	Percentage of men and women	Constant
Household (US Census)	Percentage of household size 1-7	Constant
Economic (US Census)	Median household income	Constant
Land Cover (Copernicus Sentinel-2)	NDVI, NDBI, NDMI, NDWI	Pre/Post

3.2.2 Data Processing

We followed similar data processing steps as mentioned in section 3.1.2.

- **Rasterization**: For vector data like the US census and the diseases data, we utilized vector to raster conversion tools and brought them to geotiff format.

- **Reprojection**: All of the rasters were brought to same CRS i.e., EPSG:4326 for further processing.

- **Clipping**: After identifying the Area of Interest (AOI) for as per the hurricane's landfall, we clipped the rasters based on the identified AOI. The AOI corresponding to the event is described below by the GeoJSON projected to EPSG:4326.


```
{
  "type": "Polygon",
  "coordinates": [
    [
      [-97.515061,27.796357],
      [-97.515061,28.450374],
      [-96.697953,28.450374],
      [-96.697953,27.796357],
      [-97.515061,27.796357]
    ]
  ]
}
```


- **Resampling**: Similar resampling techniques were used as mentioned in section 3.1.2.

- **Calculating Difference Channels** The difference between the land cover indices NDVI, NDMI, NDBI, NDWI, from pre and post timestamps were

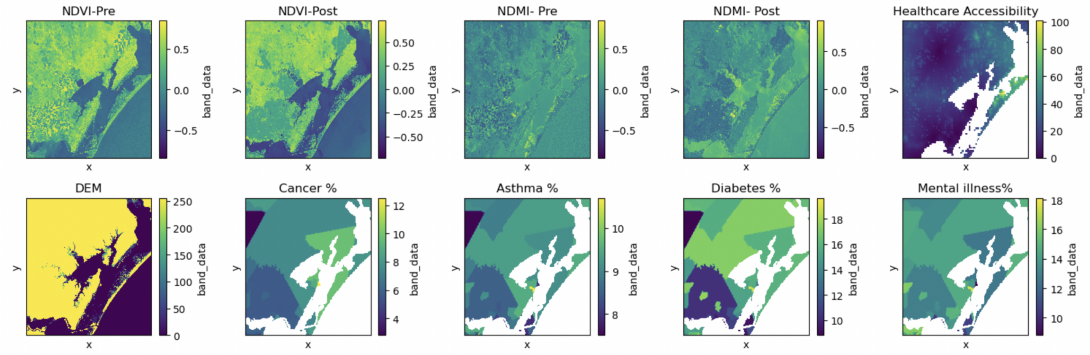

Figure 3.3: Illustration of Selected Channels from Hurricane Harvey Data Cube

calculated and added as ***Variable***_post - ***Variable***_pre. These difference data were used as proxies for damage due to the hurricane.

3.2.3 Description of Hurricane Harvey Data Cube

The data cube is stored as mentioned in section 3.1.4. The channel names and description of channels in the Hurricane Harvey data cube is shown in Table 3.4 and Figure 3.3 illustrates a subset of channels present in the data cube.

3.3 Hurricane Ida Data Cube Creation

This section discusses the process involved in creation of AI ready data cube for Hurricane Ida. Unlike for Maui wildfire and hurricane Harvey, we curate data cubes corresponding to the top 9 damaged zip codes identified and validated by NASA Impact team's analysis. We follow these steps in the curation of data cube for each zip code.

Table 3.4: Description of Different Channel in Hurricane Harvey Data Cube.

Channel Name	Description
NDVI_post	Normalized Difference Vegetation Index calculated after hurricane
NDWI_post	Normalized Difference Water Index calculated after hurricane
NDBI_post	Normalized Difference Built-up Index calculated after hurricane
NDMI_post	Normalized Difference Moisture Index calculated after hurricane
NDVI_pre	Normalized Difference Vegetation Index calculated before hurricane
NDWI_pre	Normalized Difference Water Index calculated before hurricane
NDBI_pre	Normalized Difference Built-up Index calculated before hurricane
NDMI_pre	Normalized Difference Moisture Index calculated before hurricane
NDVI_diff	Difference in Normalized Difference Vegetation Index hurricane
NDWI_diff	Difference in Normalized Difference Water Index hurricane
NDBI_diff	Difference in Normalized Difference Built-up Index hurricane
NDMI_diff	Difference in Normalized Difference Moisture Index hurricane
dem	Digital Elevation Model
H013002 - H013008	Percentage of household size 1, 2, 3, 4, 5, 6, 7+ respectively
B19013_001E	Median Household Income
P003002 - P003008	Percentages of white, black, Indian, Asian, Hawaiian, other race, and 2 or more races
P012026	Percentage of women population
P012002	Percentage of men population
bp	Percentage of population with high blood pressure
asthma	Percentage of population with asthma
cancer	Percentage of population with cancer
mental	Percentage of population with mental illness
diabetes	Percentage of population with diabetes
healthcare_accessibility	Time taken to reach nearest healthcare facility

3.3.1 Data Collection

The dataset used for the EEJ analysis was curated using the blue tarp detection on Plannetlab 3 images from NASA Impact's study, the building footprint data from Microsoft and US census data. Different categories of data collected for the data cube preparation is listed in Table 3.5.

Table 3.5: Description of different data sources and its attributes for Hurricane Ida.

Category	Attribute	Frequency
Racial (US Census)	Percentage of White, Black, African American, American Indian and Alaska Native, Asian, Native Hawaiian, Pacific Islander, Hispanic	Constant
Gender (US Census)	Percentage of men and women	Constant
Household (US Census)	Percentage of household size 1-7	Constant
Economic (US Census)	Median household income	Constant
Buildings (Microsoft) [39]	Building Footprint	Constant
Blue Tarp (Planetlab 3) [22]	Blue Tarp Detections	Multiple

3.3.2 Data Processing

We follow similar data processing as mentioned in section 3.1.2 for each zip code.

- **Rasterization**: The US census data were converted to raster data in geotiff format.
- **Re-projection**: Given the different sources of data, they come in different projection and follow different Coordinate Reference Systems (CRS). All of the raster data were brought to same CRS i.e., EPSG:4326 for further processing.
- **Clipping**: We clip the data over our identified Area of Interest. The geoJSON representing the AOI for each zip code is shown in the listing in the appendix 5.
- **Resampling**:The census raster data were resampled to match the dimensions of the blue tarp detection masks for each zip code.
- **Calculating Damage Period** For each zip code, we calculate a channel such that each pixel value is the sum of frequency of observed blue tarp occurrence over the analysis time period i.e., September 2021 through February 2022.

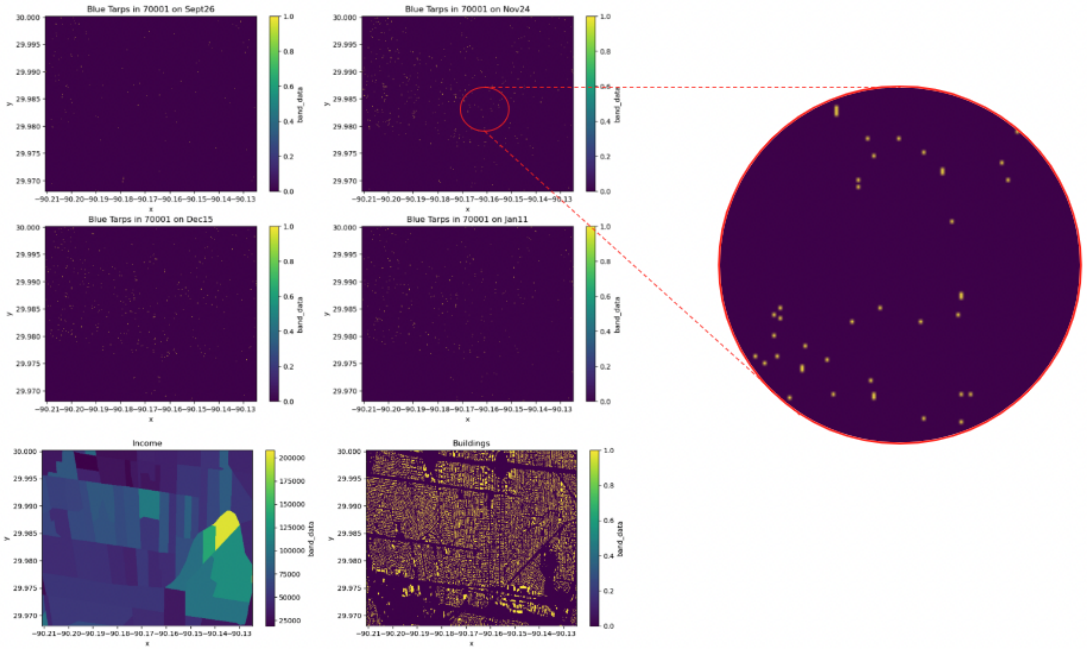

Figure 3.4: Illustration of Selected Channels in Hurricane Ida Data Cube for Zip code 70001.

3.3.3 Description of Hurricane Ida Data Cube

The data cubes corresponding to 9 zip codes are stored in NetCDF file format , with 3 dimensions : x, y and channel representing longitude, latitude and the channel name respectively in EPSG:4326 projection. The actual data is stored as data variable name *band_data* which by selectable using the channel name.

The channel names and description of the channels is in Table 3.6 and Figure 3.4 illustrates a subset of channels present in a data cube.

Table 3.6: Description of Different Channels in Ida Data Cubes.

Channel Name	Description
H013002 - H013008	Percentage of household size 1, 2, 3, 4, 5, 6, 7+ respectively
B19013_001E	Median Household Income
P003002 - P003008	Percentages of white, black, Indian, Asian, Hawaiian, other race, and 2 or more races
P012026	Percentage of women population
P012002	Percentage of men population
Sept26	Blue Tarp Detections on Sept 26, 2021
Nov24	Blue Tarp Detections on Nov 24, 2021
Dec15	Blue Tarp Detections on Dec 15, 2021
Jan11	Blue Tarp Detections on Jan 11, 2022
Feb12	Blue Tarp Detections on Feb 12, 2022
buildings	Building Footprint
damage_period	Number of occurrences of Blue tarp

3.4 Errors in the Data Cubes

As we integrated diverse data from various sources with variations in projection, resolution, and structure, tradeoffs are bound to arise. Data transformation steps such as vector to raster conversion, reprojection, resampling, and clipping are known to introduce inaccuracies in data by causing loss of information, misalignment, and compromise in quality [40]. Here we discuss some of the ways in which errors could have crept into our data generation process. Our goal is to ensure that users of these datasets and baseline models we discuss later are cognizant of the possible errors. This will allow them to take appropriate steps to mitigate the fallout from the same. Some of the ways in which inaccuracies might have entered into the data cubes are as follows.

- The **vector to raster transformation** of census data, common in all three case studies, holds the potential to introduce errors into the system. Before vector to raster transformation, it is required to specify the resolution of our raster data. If this resolution mismatches the inherent geometry of the vector data, it may fail to capture complex vector features adequately.

Similarly, geometries like polygons may possess complex curves that could get simplified during rasterization to fit the grid. As vector data is continuous, inaccuracies in the boundary values of the geometries may arise during rasterization. Therefore, the quality of the rasters depends on the complexities of vector geometry, grid resolution, boundary pixels handling, and the rasterization algorithm [41].

- **Resampling** and **reprojection** are additional data transformations applied to the channels in the datacubes. These transformations might result in spatial distortions, scaling, skewness, rotation, and other spatial distortions of the data because the data will be mapped to a different size grid. Additionally, resampling, whether up-sampling or down-sampling, employs interpolation and aggregation functions, leading to loss of information [42].

- Raster data **clipping** based on geometry is also prone to errors. Errors could be introduced due to boundary conditions, where the complex geometry edges could lie partially within a pixel, depending on the resolution and clipping algorithm used.

The errors mentioned in the datacubes have the potential to propagate into subsequent data analytics tasks. Users should remain mindful of these errors and their possible influence on downstream tasks. It's crucial for users to consult literature sources such as [43, 44, 45, 46], which offer insights and approaches for characterizing and addressing such errors. Furthermore, the results should be interpreted with the error bounds in mind.

Chapter 4. Analysis

In this section, we discuss the methods for conducting preliminary analysis on the datasets cretaed in Chapter 3 along with presenting the results. We perform in depth EEJ analysis with diverse perspectives for each of our case studies. We use statistical analysis, machine learning, image processing techniques on the datacubes curated in the previous chapter. However, it is important to note that data used for all the downstream tasks comes from the datacubes that is prone to data inaccuracies and errors as described in 3.4. These error might have propagated to all the analysis and experiments we performed [47].

4.1 Maui Wildfire Analysis Results

Following our objective to analyze environmental injustice brought by the wildfire, we sought to investigate whether disparities exist among racial, income, gender, and household size groups in the damage caused by the wildfire. We take the decrease in NDVI, NDMI, NDBI and night time light data and increase in Aerosol Index data as proxies for damage caused by the wildfire. Leveraging interpretable machine learning to we investigated if the racial, economic, gender and household features play a role in predicting the damage caused. As tree based ML algorithms are known to be interpretable and capture the non-linearity

in data, we use a Random Forest Regressor model to predict damage (in terms of NDVI, NDMI, night time light and air quality) with predictor variables as fire proximity, and the census variables. The feature importances of predictor variables signify the extent to which variations in them contribute to predicted variable changes. It's important to note that our objective is not to predict damage but rather to discern how disparities in damage are influenced by input features. The experimental setup was as follows.

4.1.1 Machine Learning for EJ analysis

We trained 4 Random Forest Regressor Models with census variables and fire proximity to predict the damage in NDVI, NDMI, Night time light and Aerosol Index. Table 4.1 shows the mean square error, R square ratio and explained variances of the models trained with different predicted variable. The explained

Table 4.1: Performance Metrics.

Predicted Variable	Mean Square Error	Explained Variance Ratio
NDVI damage	0.022	0.47
NDMI damage	0.005	0.32
Night time light damage	0.012	0.24
Aerosol Index damage	1.009	0.7

variance ratio is the proportion of the variance in damage variable that was captured by the predictor variables. The feature importances of input variables in

shaping disparities in the damage in variables are shown in Figure 4.1. An explained variance ratio of 0.47 for NDVI damage suggests that the census variable and fire proximity can capture 47% of the variance in NDVI damage caused by the fire. Similarly, the input features could capture 70% of variance in the air quality data, 24% of variance in night time light damage and 32% variance in the damage in terms of moisture. From the results we observed that the disparites in air quality damage is shaped in a greater extent by the racial, gender, household size, income features. We see the feature importances to see how different races, genders, income groups and household size play a role in shaping disparities in damage in vegetation due to fire. In the figure 4.1, we can see that for predicting all damage variables, fire proximity and income are identified as the top most important features. It suggests that damages in a location vary with different income groups as well as with its closeness to the wildfire. After fire proximity and income,

- $\Rightarrow$ damage in vegetation (NDVI) varies with the percentage of household size and women's population.
- $\Rightarrow$ damage in moisture (NDMI) varies with the percentage of asian population and men's population.
- $\Rightarrow$ damage in air quality (aerosol index) varies with the percentage of racial category "two or more races" and the household size.
- $\Rightarrow$ damage in economic/industrial activity (night time light) varies with the percentage of racial category "two or more races" and the household size.

Figure 4.2 illustrates the correlation between the input features with the different damage variables. It compliments the feature importance plot by statistically representing how the input variables are correlated with the damage variables. While the feature importances give us the extent of any input variable in causing disparities in damage variable, its corresponding Pearson's correlation coefficient value sheds light in either they share a positive or negative relationship. For instance, the damage in air quality varies with different income groups (from the feature importance plot) and they are negatively correlated with each other. This implies that low economic group communities suffered from relatively high damage in air quality, which could be a serious environmental injustice issue with respect to air quality and low income. Likewise, the following are a few environmental injustice issues that was inferred using similar interpretation:

- Low income group communities suffered more damage in vegetation, air quality and economic activity.
- The racial group - two or more category suffered more damage in economic activity.
- Communities with larger household size suffered more damage in economic activity.

4.1.2 Quantitative and Qualitative Analysis

In the previous section, we used statistical and machine learning based approach to infer the presence of environmental injustice. The quantitative anal-

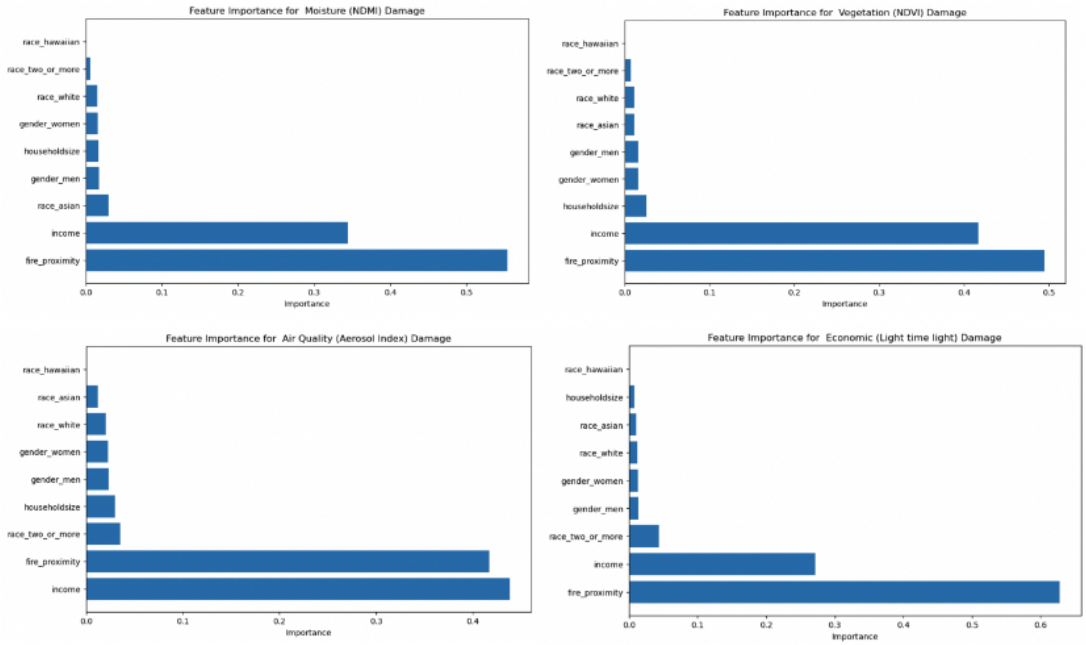

Figure 4.1: Feature Importances for Different Damage Variables.

ysis deals with quantifying the extent to which the injustice occurred and the qualitative analysis deals with the spatial information of the injustice issues.

4.1.2.1 Damage Calculation using Remote Sensing Variables

Remote sensing variables like NDVI, NDMI, aerosol index and night time light data vary temporally before and after the wildfire. The difference in the pre and post timestamps of these variables can be used to quantify damage brought by the wildfire. Figure 4.3 illustrates the percentage of damage in pixels over the total pixels covering Maui Island. There had been **80%** damage in NDVI, and **70%** of damage has been observed in close proximity to the fire. Similarly, there has been **16%** damage in night time lights, and **13%** damage had been observed

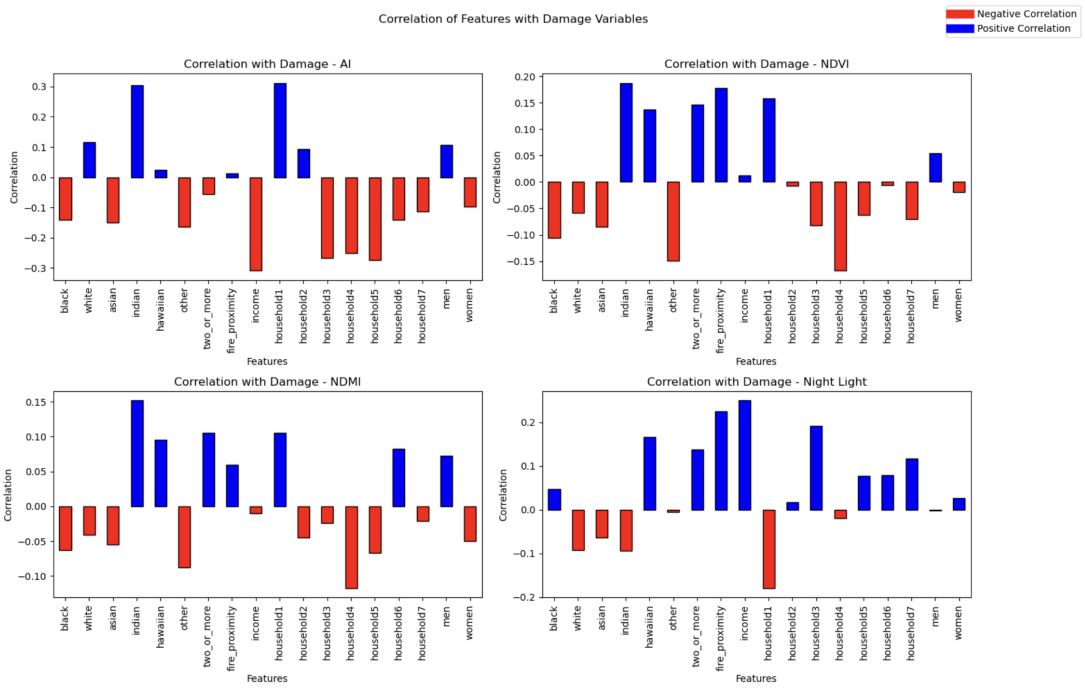

Figure 4.2: Correlation of input variables with damage variables.

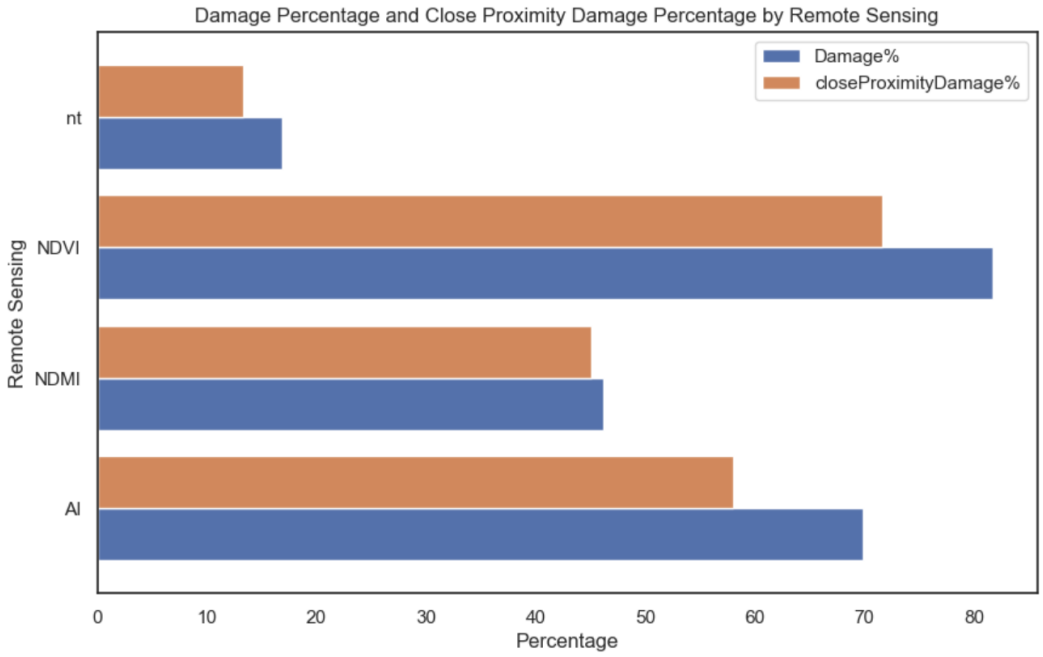

Figure 4.3: Damage Percentage in Remote Sensing Variables in Maui Island and in Close Proximity to Fire.

in close proximity of the fire. The damage plot shows that the effects of wildfire was depicted more by the difference of NDVI, Aerosol Index while night time light data is less affected by the wildfire. This discrepancy may arise due to the direct impact of a fire event on vegetation damage and air quality degradation, whereas decrease in night light data may occur as an indirect consequence.

4.1.2.2 Damage Analysis with respect to Census Variables

Since environmental injustice is a very subjective term and is always with respect to certain community, we analyse different combinations of damage variables paired with a census variable and the fire proximity. We generated multiple geo-referenced 3 channeled tiff files such that the first channel is any of the differ-

ence in remote sensing variable (NDVI_diff, NDMI_diff, AI_diff etc), the second channel is any of the census variables (H013002, P012026 etc) and the third channel is the fire proximity. To generalize, each file has a channel representing a damage variable, a EJ variable and a channel corresponding to the fire. For each of such images, we apply the following thresholding conditions and create a mask from the image satisfying these conditions:

- **Difference Variable** $<$ mean (for NDVI, NDMI, Night time light)
  $>$ mean (for Aerosol Index)
- **EEJ Variable** $>$ mean
- **Fire Proximity** $>$ mean

Each mask will have the injustice mask for that particular EEJ variable based on the damage from that particular remote sensing data due to the wildfire. We saved these images and its corresponding masks as the filenames DDD_CCC_FFF.tiff and DDD_CCC_FFF_mask.tiff respectively. Figure 3.5 shows the summary for quantitative analysis of each EEJ mask giving out the percentage of area identified as injustice area over the total damaged area. For each image, the percentages of injustice area due to wildfire with respect to a particular damage variable and EEJ variable is calculated as shown in the bar charts of Figure 4.4. It quantifies the environmental injustice mask associated to each damage variable and EEJ variable as percentages. And the qualitative analysis representing the mask plots, show exactly where the injustice cases are spatially. For instance, **11%** of the pixels damaged in terms of night light have suffered

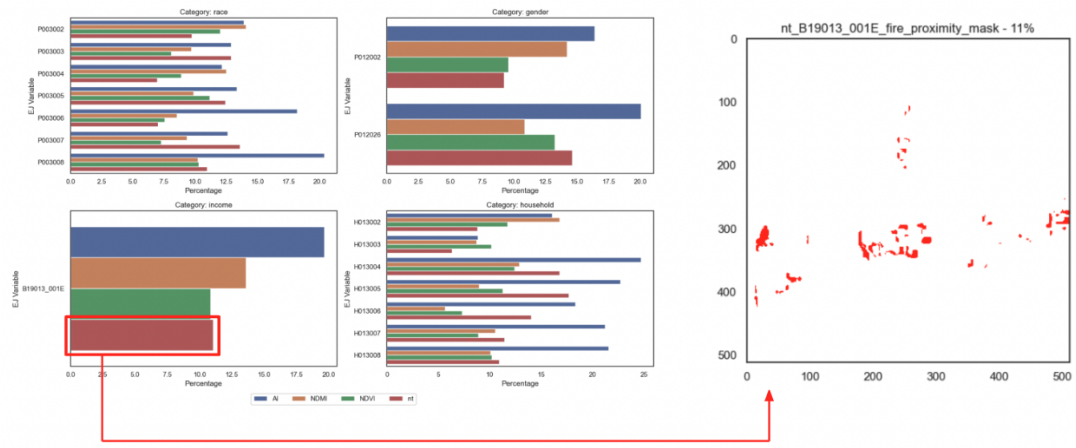

Figure 4.4: Quantitative and Qualitative Analysis of EEJ.

environmental injustice with respect to income groups and the corresponding ***nt_B19013_001E_fire_proximity_mask.tiff*** shows the spatial distribution of the injustice area.

4.1.3 EEJ Masks Creation

With the goal of identifying environmental injustice issues, we present diverse methods for creating masks to highlight areas of environmental injustice. Our approaches include thresholding techniques, unsupervised learning methods, and leveraging inference from a foundation model, all of which are discussed below.

4.1.3.1 Using Thresholding for EEJ Masks Generation

Thresholding involves applying specific conditions to an image, generating a mask based on the satisfying conditions. Widely utilized in image processing,

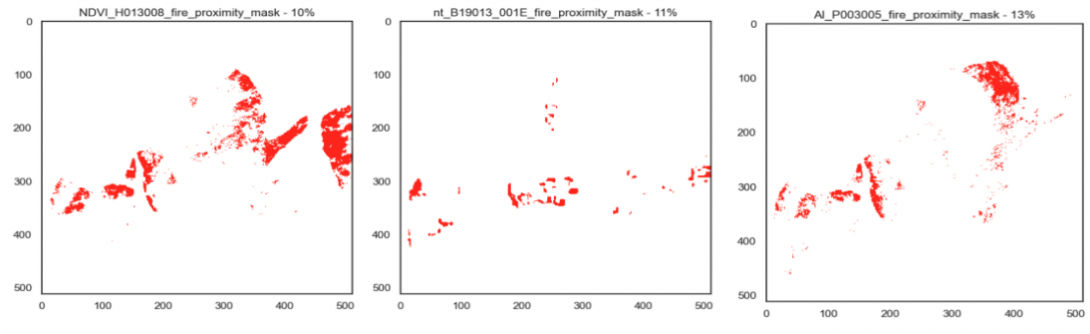

Figure 4.5: EEJ Masks Creation Using Thresholding.

thresholding is a fundamental technique for image segmentation, dataset generation, and internal steps in deep learning algorithms. As the concept of curating Environmental Equity and Justice (EEJ) masks is novel, we propose an initial process based on our conception of environmental injustice. We define an environmental injustice area as one that experiences more damage during an event and exhibits increased vulnerability in housing, income, race, or gender. Applying the thresholding conditions outlined in Section 3.3.2, we create EEJ masks as a preliminary step in dataset creation for the EEJ domain. Due to the complexity and sensitivity of EEJ data, crowd-sourcing is impractical, leading us to rely on a basic understanding of environmental injustice to shape the masks. This approach serves as a foundational step for EEJ dataset creation and provides a ground truth for subsequent tasks. Figure 4.5 depicts instances of masks generated using image thresholding.

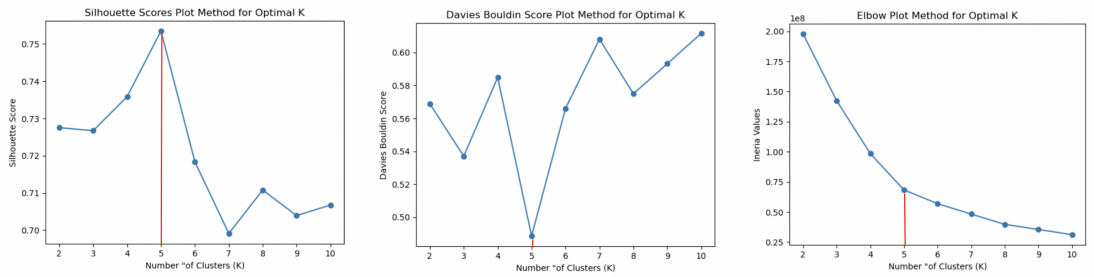

Figure 4.6: Comparing Performance Metrics to find the Optimal Number of Clusters in Data.

4.1.3.2 Using Unsupervised Learning for EEJ Masks Generation

We used K-means clustering algorithm to segment out clusters based on the values of the three channels of the image. Before feeding images to K-means algorithm, we normalized each channel from 0-255. We found the optimal number of clusters to be 5 as illustrated by the Silhoutte score, Davies Bouldin score and Elbow plot in Figure 4.6. Figure 4.9 shows an example of mask generated using k-means. Figure 4.7 shows the distribution of the maximum percentage of overlapping pixels between a k-means segment and ground truth for all the images. The range of percentage overlap lies with in the range of 40% to a 100%.

4.1.3.3 Using Foundation Model for EEJ Masks Generation

Foundation models are pre-trained on a large amount of data and have the state of the art performance. We utilized the Segment Anything Model (SAM) to segment the images and Figure 4.9 shows an example of the segmented image. On average, it was able to segment out 28 segments in the input images. Figure 4.8 shows the distribution of the maximum percentage of overlapping pixels between

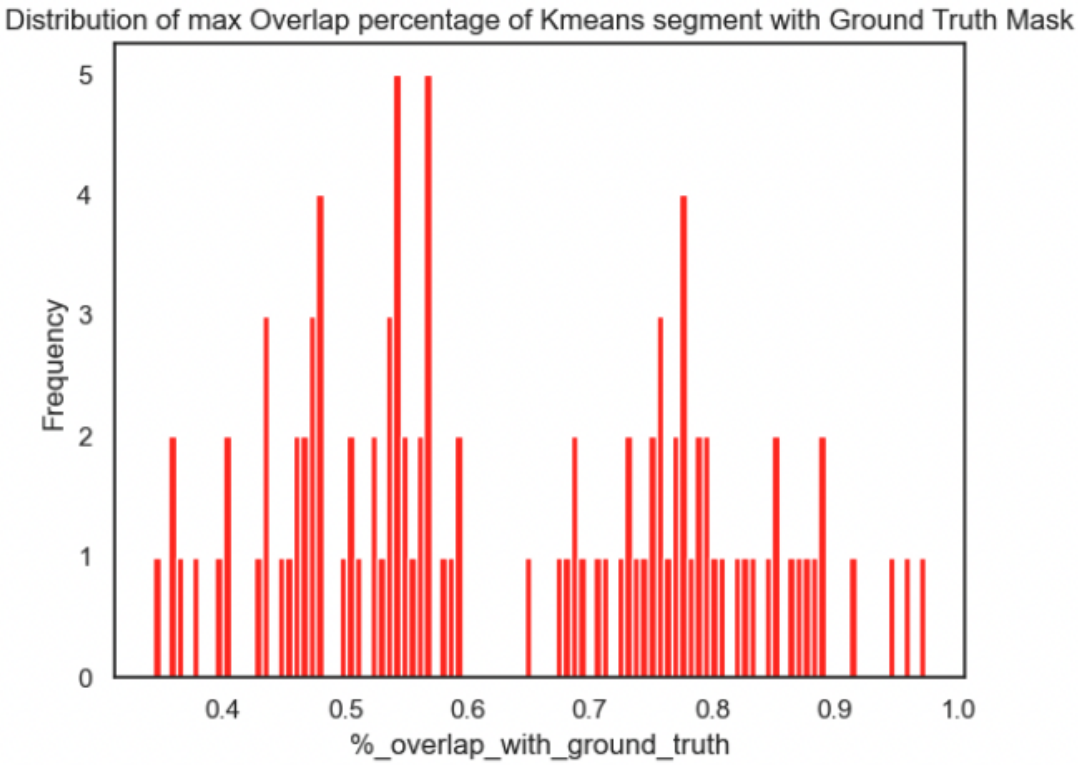

Figure 4.7: Intersection Between K means cluster and Thresholded Mask.

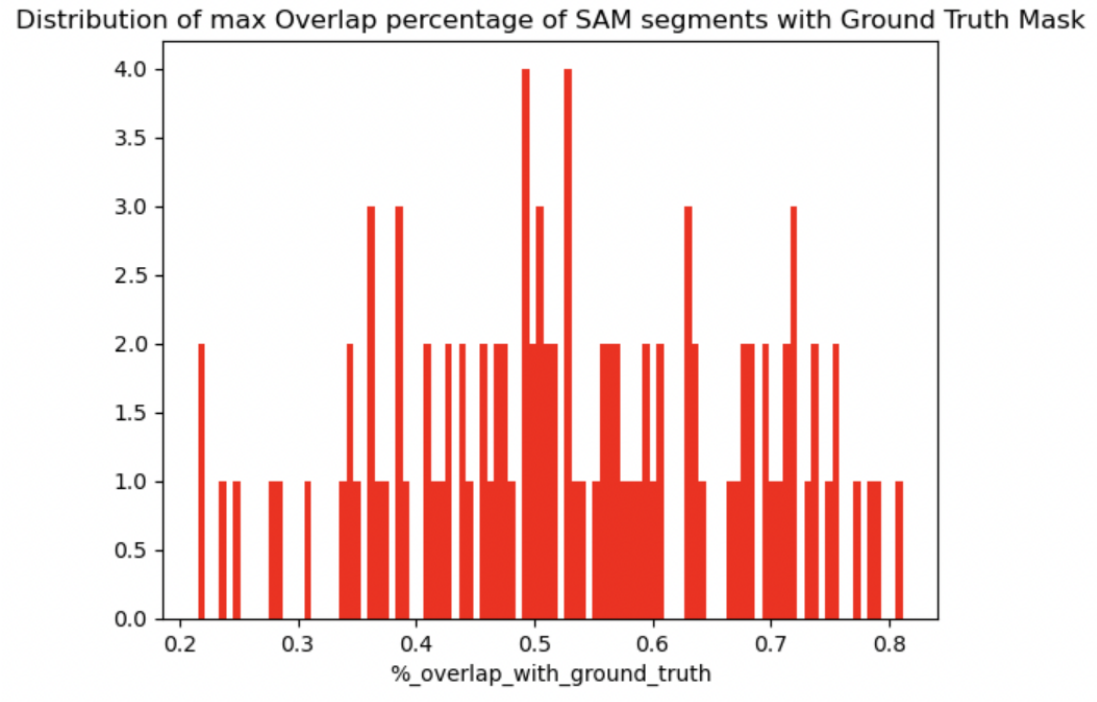

Figure 4.8: Intersection Between SAM cluster and Thresholded Mask.

a SAM segment and ground truth for all the images. The range of percentage overlap lies with in the range of 20% to 80%.

4.1.4 Comparison of EEJ Masks

Figure 4.9 shows the comparison of the three techniques used for segmentation. If we compare the masks generated through these processes, we can see overlaps between the masks from thresholding and k-mmeans clustering, suggesting that K means can be a alternative method to generate the masks. However, the masks generated using SAM inference are very granular and have less overlap with our thresholded masks, evident in Figure 4.9 . In future, it would be

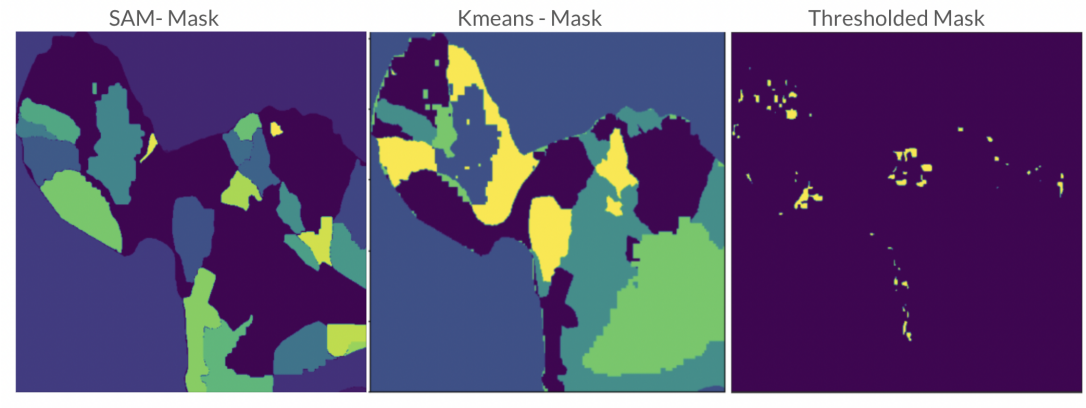

Figure 4.9: Comparision of Masks Genrated using SAM, Kmeans Clustering and Thresholding.

a worthwhile experiment to label more images using k-means and fine tune the SAM model on the labeled data and see the output.

4.2 Hurricane Harvey Analysis

We tailored our analysis of Hurricane Harvey to focus on its impact on health-vulnerable communities, primarily justified by the flooding it caused [48]. Studies have shown that the occurrence of Hurricane Harvey led to increased physical and mental health needs [49]. A significant aspect of our analysis aimed to understand how individuals with chronic diseases were affected by flooding in relation to the availability of healthcare services. We used proxies such as NDMI, NDWI, and NDVI to identify flooded areas, as these indices are considered effective for flood detection [21]. Additionally, we took the topography of the area into consideration, recognizing its importance in flood analysis [50]. Our explo-

ration sought to uncover relationships among disease data, healthcare availability, topography, and the damage caused by flooding.

4.2.1 Relationship Among Diseases and Healthcare Accessibility

We examined the correlations between the prevalence of chronic health conditions—such as high blood pressure, cancer, mental illness, diabetes, and asthma—and the availability of healthcare, measured in terms of the time required to reach health facilities. Figure 4.10 reveals strong positive correlations between the population affected by diseases and the time it takes to access healthcare. This suggests potential disparities in healthcare accessibility times for populations vulnerable to health issues.

4.2.2 Relationship of Flooding with Health Variables

We utilized a Random Forest Regressor Model incorporating disease variables, healthcare accessibility, and topography data to predict flooding damages. The observed mean square error was **0.01**, with an explained variance ratio of **36%**. This indicates that 36% of the variance in flood damage can be explained by our predictor variables. Figure 4.11 illustrates the feature importances of these predictor variables, highlighting healthcare accessibility time and topography (DEM) as the top contributing factors shaping disparities in flood damage. The Pearson correlation coefficient in Figure 4.11 provides insights into the direction of the relationship between health and topographical variables with damage. Areas with low elevation show a negative correlation with flooding, while health-

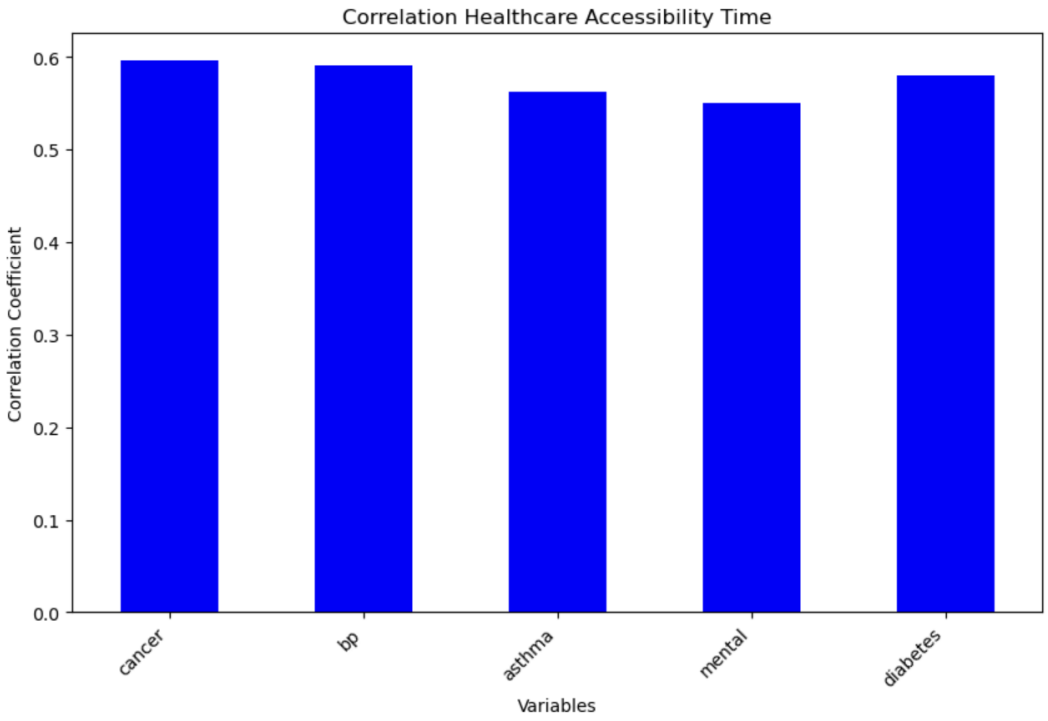

Figure 4.10: Correlation of Disease Variables with Healthcare Accessibility.

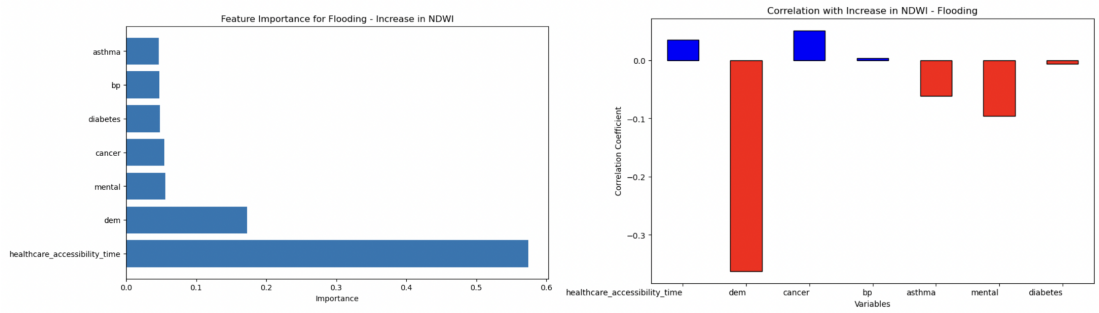

Figure 4.11: Feature Importances of Input variables in Random Forest Regressor and Correlation Plot of Input Variables with Increase in NDWI.

care accessibility time, population with cancer, and blood pressure exhibit positive correlations. These analyses provide valuable insights into environmental justice issues related to health and disasters.

4.3 Hurricane Ida Analysis

As discussed in the previous chapter, we generated analysis-ready data cubes for each zip code, encompassing building footprints, time series blue tarp detections, and demographic and socioeconomic data. This section focuses on analyzing the characteristics of zip codes exhibiting slow and fast recovery. The line plot presented in Figure 4.12 visualizes blue tarp detections over building footprints across time, providing insights into the recovery processes of different zip codes. The plot specifically showcases the top 9 most damaged zip codes identified by the NASA IMPACT team. Additionally, we introduced a metric called "recovery_rate" to quantify the recovery in each zip code. A higher recovery_rate indicates a faster recovery from the maximum damaged condition to the last observed state on February 12, 2022. As depicted in Figure 4.13, zip code

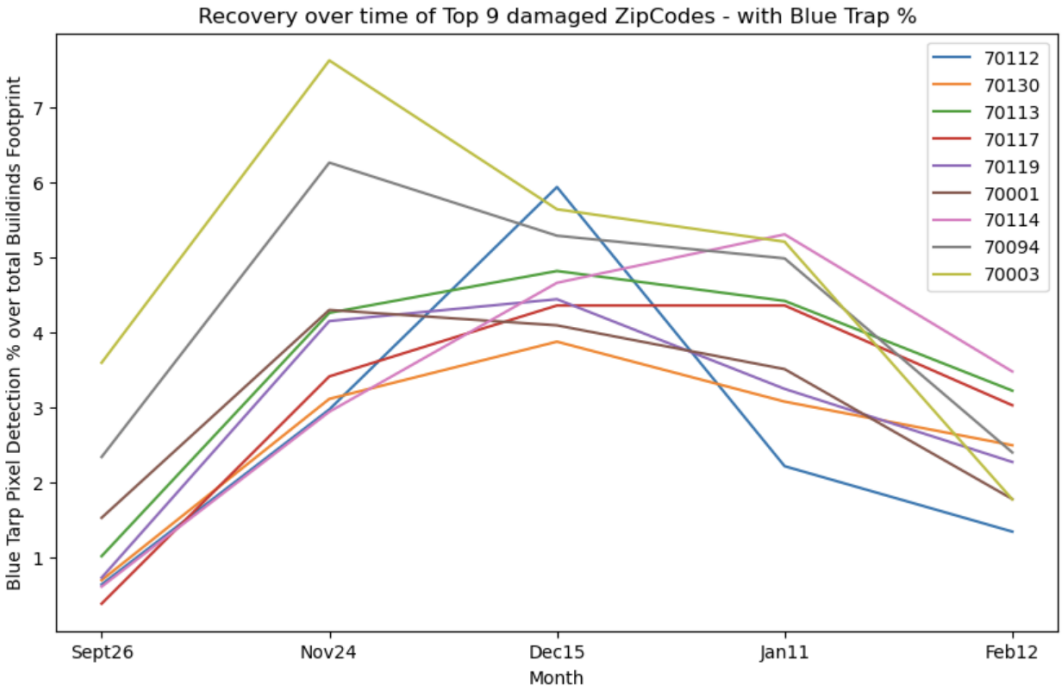

Figure 4.12: Time Series Plot of Blue Tarp Detections as Percentage of Building Footprints for Different Zip Codes.

70112 exhibits the highest recovery rate, suggesting a significant decrease in blue tarps within a short time period. In contrast, zip codes such as 70130, 70117, and 70001 show slower recovery rates. To explore the correlation between socioeconomic and demographic features and the recovery rate of different zip codes, Figure 4.14 illustrates these relationships. Our observations reveal that zip codes with a higher population of Asian residents tend to recover faster, while those with a larger population of "two or more" races recover more slowly. Additionally, there is an indication that zip codes with higher income levels experience a higher recovery rate.

For a more detailed analysis, we created a new raster named "damage_period" from the time series of blue tarp detections. This raster assigns values from 0 to 5, where 0 indicates that the pixel has never been detected as a blue tarp, and 5 indicates that the pixel has been consistently identified as a blue tarp in all five timestamps. The "damage_period" metric provides a pixel-level estimate of the damage period.

Figure 4.15 utilizes the "damage_period" channel to visualize the damage periods in different zip codes. A value of 5 in a pixel suggests prolonged damage, while lower values indicate varying degrees of damage observation. When comparing zip codes like 70003 with 70001, it is evident that 70003 has a higher number of pixels observed as blue tarps once but fewer pixels observed consistently across all months. On the other hand, zip codes like 70001 and 70114 exhibit similar observations of pixels seen throughout all months, suggesting potential recovery issues in these areas.

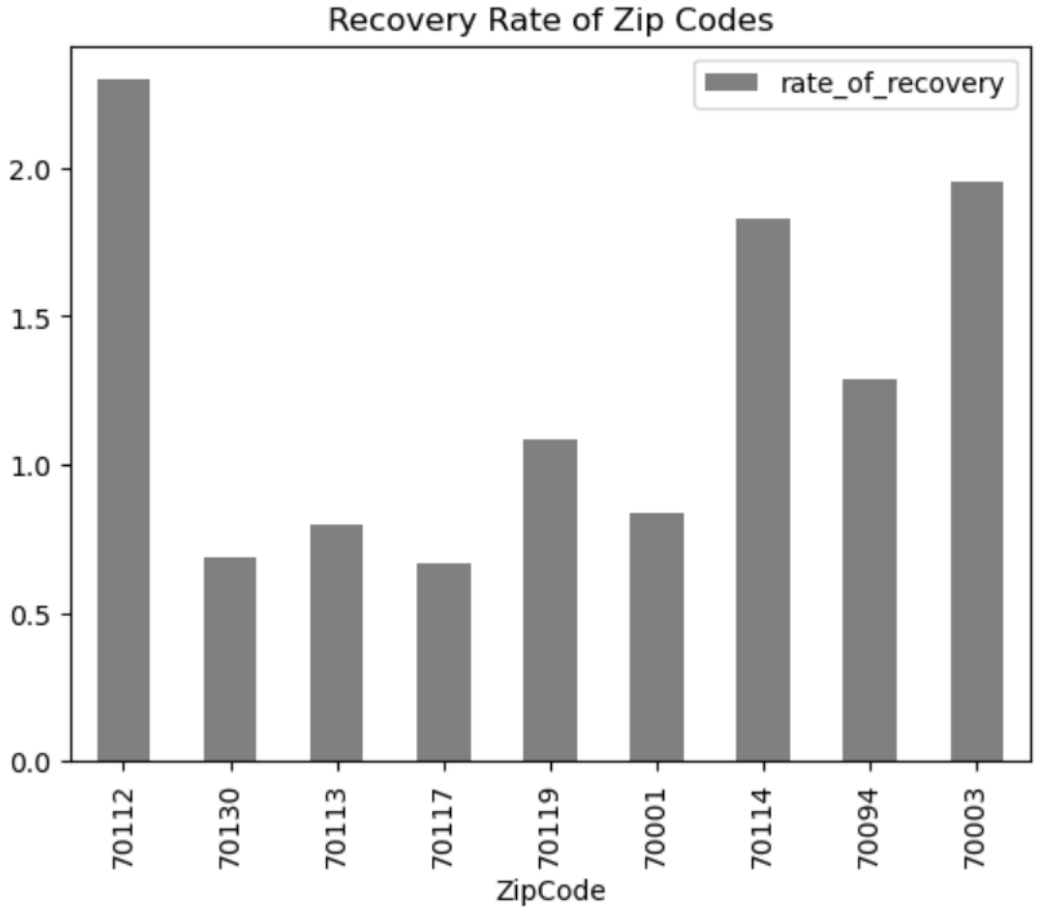

Figure 4.13: Recovery Rates for Different Zip Codes.

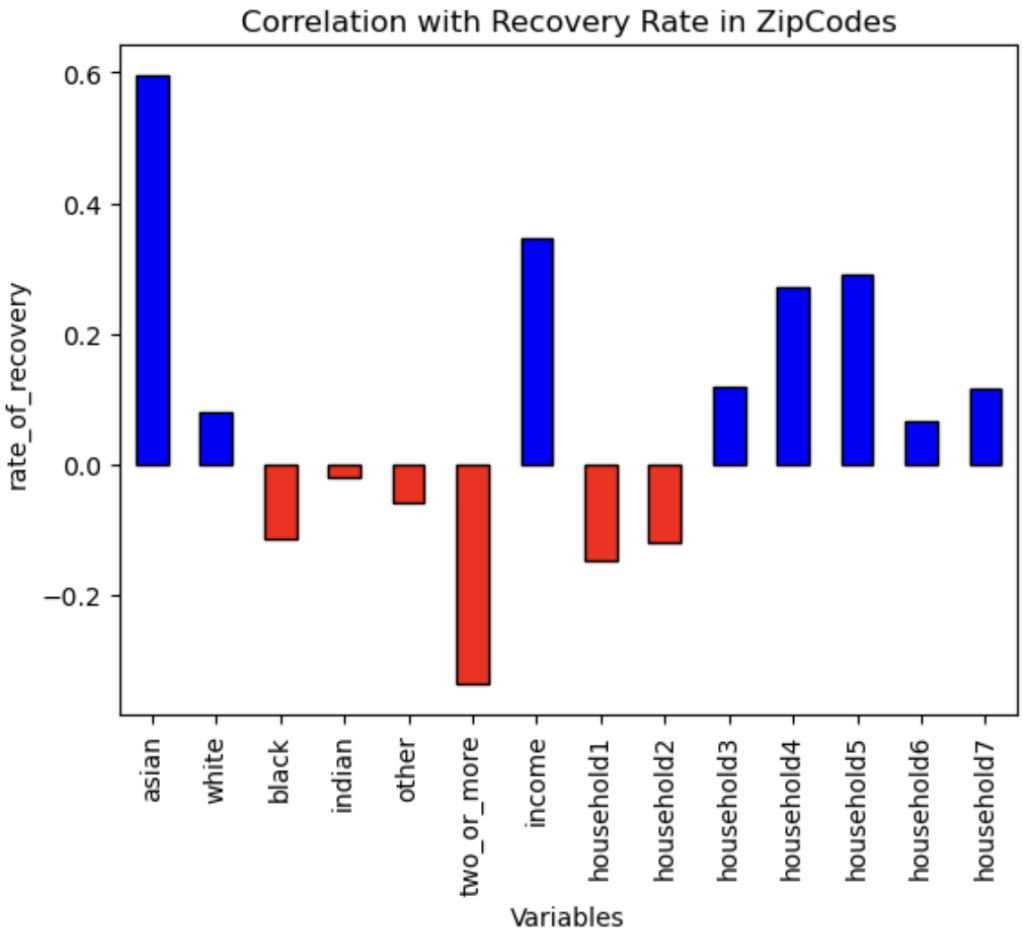

Figure 4.14: Correlation Plot among Socioeconomic Variables with Recovery Rate.

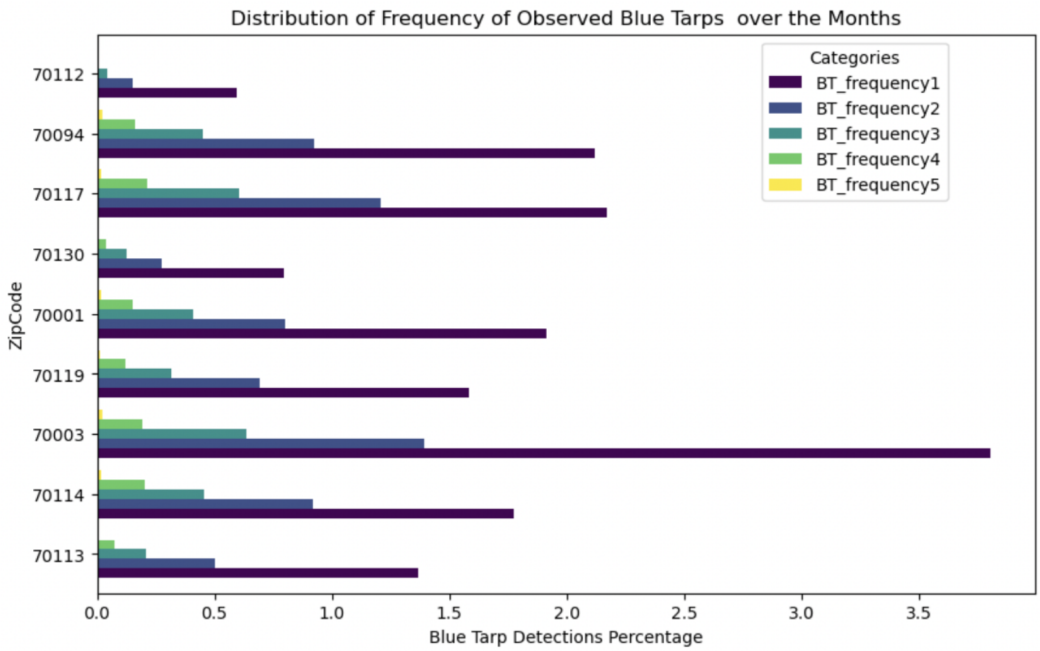

Figure 4.15: Distribution of Frequency of Blue Tarp Occurrences for Different Zip Codes.

Chapter 5. Conclusion and Future Works

In conclusion, our analysis of disasters—Maui Wildfire, Hurricane Harvey, and Hurricane Ida—provide nuanced insights into environmental impact and socioeconomic vulnerabilities. Leveraging interpretable machine learning, image processing techniques and statistical methods, we illustrate how socioeconomic factors shape post-disaster disparities in terms of damage and recovery. Our innovative methods, including EEJ mask creation and foundation model utilization, contribute to a deeper understanding of Equity and Environmental Justice issues in disaster contexts.

The future avenues for this work include incorporating additional attributes. Furthermore, we anticipate augmenting the dataset with information related to a broader spectrum of events, such as pandemics and wars, thereby establishing a comprehensive data repository. This dataset will lay the groundwork for diverse machine learning applications, empowering researchers to employ data-centric AI methodologies, including data augmentation and confident learning [51]. We also anticipate working on algorithms for quantifying as well as minimizing the error introduced during the data transformation process. We envision utilizing foundational models and fine-tuning them after we have collected enriched data on many events [52]. Additionally, we envision making the dataset accessible through

community-contributed platforms such as Hugging Face and Zenodo in a crossiant format [53]. We hope this research will open new avenues in the research toward Equity and Environmental justice.

References

[1] Francis O Adeola. *Industrial disasters, toxic waste, and community impact: health effects and environmental justice struggles around the globe*. Lexington Books, 2012.

[2] Tatyana G Krupnova, Olga V Rakova, Kirill A Bondarenko, and Valeria D Tretyakova. Environmental justice and the use of artificial intelligence in urban air pollution monitoring. *Big Data and Cognitive Computing*, 6(3):75, 2022.

[3] Hawaii wilffires timeline maui. `https://www.cnn.com/interactive/2023/08/hawaii-wildfires-timeline-maui-lahaina-dg/index.html`. Accessed: 2024-02-10.

[4] Maui unemployment rate. `https://www.mauinews.com/news/local-news/2023/11/mauis-unemployment-rate-shot-up-to-8-4-in-september/`. Accessed: 2024-02-10.

[5] Toni Sebastian, Kasper Lendering, Baukje Kothuis, Nikki Brand, Bas Jonkman, Pieter van Gelder, Maartje Godfroij, Bas Kolen, Tina Comes, Stef Lhermitte, Kenny Meesters, Bartel van de Walle, Amir Ebrahimi Fard, Scott Cunningham, N. Khakzad, and Vittorio Nespeca. *Hurricane Harvey Report: A fact-finding effort in the direct aftermath of Hurricane Harvey in the Greater Houston Region*. Delft University Publishers, 2017.

[6] Hurricane ida. `https://www.nhc.noaa.gov/data/tcr/AL092021_Ida.pdf`. Accessed: 2024-02-10.

[7] Darius Phiri, Matamyo Simwanda, Serajis Salekin, Vincent R Nyirenda, Yuji Murayama, and Manjula Ranagalage. Sentinel-2 data for land cover/use mapping: A review. *Remote Sensing*, 12(14):2291, 2020.

[8] Vladimir Tabunschik, Roman Gorbunov, and Tatiana Gorbunova. Unveiling air pollution in crimean mountain rivers: Analysis of sentinel-5 satellite images using google earth engine (gee). *Remote Sensing*, 15(13):3364, 2023.

[9] Wilfrid Schroeder, Patricia Oliva, Louis Giglio, and Ivan A Csiszar. The new viirs 375 m active fire detection data product: Algorithm description and initial assessment. *Remote Sensing of Environment*, 143:85–96, 2014.

[10] U.S. Census Bureau. 2010 census. U.S. Census Bureau, 2010.

[11] The White House. Justice40 initiative.

[12] A Myrick Freeman III. Distribution of environmental quality. In *The Economic Approach to Environmental Policy*, pages 72–107. Edward Elgar Publishing, 1998.

[13] William Bowen. An analytical review of environmental justice research: what do we really know? *Environmental management*, 29:3–15, 2002.

[14] Stephan Rasp, Peter D Dueben, Sebastian Scher, Jonathan A Weyn, Soukayna Mouatadid, and Nils Thuerey. Weatherbench: a benchmark data set for data-driven weather forecasting. *Journal of Advances in Modeling Earth Systems*, 12(11):e2020MS002203, 2020.

[15] Clara Betancourt, Timo Stomberg, Ribana Roscher, Martin G Schultz, and Scarlet Stadtler. Aq-bench: A benchmark dataset for machine learning on global air quality metrics. *Earth System Science Data*, 13(6):3013–3033, 2021.

[16] Yu-Hsuan Ho, Zhewei Liu, Cheng-Chun Lee, and Ali Mostafavi. Ml4ej: Decoding the role of urban features in shaping environmental injustice using interpretable machine learning. *arXiv preprint arXiv:2310.02476*, 2023.

[17] Georgiy Bobashev, Ignacio Segovia-Dominguez, Yulia R Gel, James Rineer, Sarah Rhea, and Hui Sui. Geospatial forecasting of covid-19 spread and risk of reaching hospital capacity. *SIGSPATIAL Special*, 12(2):25–32, 2020.

[18] Greater Impact. How disasters affect people of low socioeconomic status. *Substance Abuse and Mental Health Services Administration (SAMHSA). Disaster Technical Assistance Center Supplemental Research Bulletin*, 2017.

[19] Jayajit Chakraborty, Sara E Grineski, and Timothy W Collins. Hurricane harvey and people with disabilities: Disproportionate exposure to flooding in houston, texas. *Social Science & Medicine*, 226:176–181, 2019.

[20] Sangung Park, Tong Yao, and Satish V Ukkusuri. Spatiotemporal heterogeneity reveals urban-rural differences in post-disaster recovery. *npj Urban Sustainability*, 4(1):2, 2024.

[21] E Stoyanova. Remote sensing for flood inundation mapping using various processing methods with sentinel-1 and sentinel-2. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 48:339–346, 2023.

[22] NASA Earthdata. Hurricane Maria and Ida Dashboard.

[23] Marie-Therese Puth, Markus Neuhäuser, and Graeme D Ruxton. Effective use of spearman's and kendall's correlation coefficients for association between two measured traits. *Animal Behaviour*, 102:77–84, 2015.

[24] Christian Robert. *Machine learning, a probabilistic perspective*. Taylor & Francis, 2014.

[25] Lingjian Yang, Songsong Liu, Sophia Tsoka, and Lazaros G Papageorgiou. A regression tree approach using mathematical programming. *Expert Systems with Applications*, 78:347–357, 2017.

[26] Shi Na, Liu Xumin, and Guan Yong. Research on k-means clustering algorithm: An improved k-means clustering algorithm. In *2010 Third International Symposium on intelligent information technology and security informatics*, pages 63–67. Ieee, 2010.

[27] MFM Yunoh, S Abdullah, MHM Saad, ZM Nopiah, and MZ Nuawi. Fatigue feature extraction analysis based on a k-means clustering approach. *Journal of Mechanical Engineering and Sciences*, 8:1275–1282, 2015.

[28] Muhammad Shaheen, Saeed Iqbal, and Fazl e Basit. Labeled clustering a unique method to label unsupervised classes. In *8th International Conference for Internet Technology and Secured Transactions (ICITST-2013)*, pages 210–214, 2013.

[29] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

[30] Christian Mattjie, Luis Vinicius de Moura, Rafaela Cappelari Ravazio, Lucas Silveira Kupssinskü, Otávio Parraga, Marcelo Mussi Delucis, and Rodrigo Coelho Barros. Zero-shot performance of the segment anything model (sam) in 2d medical imaging: A comprehensive evaluation and practical guidelines, 2023.

[31] Yizhe Zhang, Tao Zhou, Shuo Wang, Ye Wu, Pengfei Gu, and Danny Z. Chen. Samdsk: Combining segment anything model with domain-specific knowledge for semi-supervised learning in medical image segmentation, 2023.

[32] Daochen Zha, Zaid Pervaiz Bhat, Kwei-Herng Lai, Fan Yang, Zhimeng Jiang, Shaochen Zhong, and Xia Hu. Data-centric artificial intelligence: A survey. *arXiv preprint arXiv:2303.10158*, 2023.

[33] Johannes Jakubik, Sujit Roy, CE Phillips, Paolo Fraccaro, Denys Godwin, Bianca Zadrozny, Daniela Szwarcman, Carlos Gomes, Gabby Nyirjesy, Blair Edwards, et al. Foundation models for generalist geospatial artificial intelligence. *arXiv preprint arXiv:2310.18660*, 2023.

[34] Manish Kumar, RB Singh, Anju Singh, Ram Pravesh, Syed Irtiza Majid, and Akash Tiwari. Referencing and coordinate systems in gis. In *Geographic Information Systems in Urban Planning and Management*, pages 25–46. Springer, 2023.

[35] Sk Sazid Mahammad and R Ramakrishnan. Geotiff-a standard image file format for gis applications. *Map India*, pages 28–31, 2003.

[36] R. Rew and G. Davis. Netcdf: an interface for scientific data access. *IEEE Computer Graphics and Applications*, 10(4):76–82, 1990.

[37] DJ Weiss, Andy Nelson, CA Vargas-Ruiz, K Gligorić, S Bavadekar, Evgeniy Gabrilovich, A Bertozzi-Villa, J Rozier, HS Gibson, T Shekel, et al. Global maps of travel time to healthcare facilities. *Nature medicine*, 26(12):1835–1838, 2020.

[38] Cdc environmental justice index. `https://www.atsdr.cdc.gov/placeandhealth/eji/index.html`.

[39] Microsoft. USBuildingFootprints. `https://github.com/microsoft/USBuildingFootprints`.

[40] Zachary J Christman and John Rogan. Error propagation in raster data integration. *Photogrammetric Engineering & Remote Sensing*, 78(6):617–624, 2012.

[41] Arnold Bregt, J. DENNEBOOM, and Huynh Van Y. Determination of rasterizing error: a case study with the soil map of the netherlands. *International Journal of Geographical Information Science*, 5:361–367, 07 1991.

[42] A Jon Kimerling. Predicting data loss and duplication when resampling from equal-angle grids. *Cartography and Geographic Information Science*, 29(2):111–126, 2002.

[43] Joe G Greener, Shaun M Kandathil, Lewis Moffat, and David T Jones. A guide to machine learning for biologists. *Nature reviews Molecular cell biology*, 23(1):40–55, 2022.

[44] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, et al. *An introduction to statistical learning*, volume 112. Springer, 2013.

[45] Vladimir Naumovich Vapnik, Vlamimir Vapnik, et al. Statistical learning theory. 1998.

[46] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[47] Arthur Elmes, Hamed Alemohammad, Ryan Avery, Kelly Caylor, J Ronald Eastman, Lewis Fishgold, Mark A Friedl, Meha Jain, Divyani Kohli, Juan Carlos Laso Bayas, et al. Accounting for training data error in machine learning applied to earth observations. *Remote Sensing*, 12(6):1034, 2020.

[48] Kimberly A Chambers, Irfan Husain, Yashwant Chathampally, Alan Vierling, Marylou Cardenas-Turanzas, Fanni Cardenas, Kunal Sharma, Samuel Prater, and Jonathan Rogg. Impact of hurricane harvey on healthcare utilization and emergency department operations. *Western journal of emergency medicine*, 21(3):586, 2020.

[49] Robert Bozick. The effects of hurricane harvey on the physical and mental health of adults in houston. *Health & Place*, 72:102697, 2021.

[50] Salvatore Manfreda, Margherita Di Leo, and Aurelia Sole. Detection of flood-prone areas using digital elevation models. *Journal of Hydrologic Engineering*, 16(10):781–790, 2011.

[51] Curtis Northcutt, Lu Jiang, and Isaac Chuang. Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research*, 70:1373–1411, 2021.

[52] Gengchen Mai, Chris Cundy, Kristy Choi, Yingjie Hu, Ni Lao, and Stefano Ermon. Towards a foundation model for geospatial artificial intelligence (vision paper). In *Proceedings of the 30th International Conference on Advances in Geographic Information Systems*, pages 1–4, 2022.

[53] MLCommons. Croissant: A framework for mlcommons datasets. `https://github.com/mlcommons/croissant`.

Appendix A. Hurricane Ida Zip Codes Area of Interests

{‘ZipCode’: ‘70113’, ‘type’: ‘Polygon’,

‘coordinates’: [[[-90.09445795732412,29.93161641935578],

[-90.09445795732412, 29.955384714775057],

[-90.0718512703276, 29.955384714775057],

[-90.0718512703276, 29.93161641935578],

[-90.09445795732412, 29.93161641935578]]]}

{‘ZipCode’: ‘70114’, ‘type’: ‘Polygon’,

‘coordinates’: [[[-90.0586863748698, 29.907222642478107],

[-90.0586863748698, 29.95839298274166],

[-90.02970573396384, 29.95839298274166],

[-90.02970573396384, 29.907222642478107],

[-90.0586863748698, 29.907222642478107]]]}

{‘ZipCode’: ‘70003’, ‘type’: ‘Polygon’,

‘coordinates’: [[[-90.23808041410577, 29.973464107406013],

[-90.23808041410577, 30.03586332908945],

[-90.18920350585636, 30.03586332908945],

[-90.18920350585636, 29.973464107406013],

[-90.23808041410577, 29.973464107406013]]]}

{‘ZipCode’: ‘70119’, ‘type’: ‘Polygon’,

‘coordinates’: [[[-90.1140861611854, 29.95312106759227],

[-90.1140861611854, 29.992854032616474],
[-90.05812046307409, 29.992854032616474],
[-90.05812046307409, 29.95312106759227],
[-90.1140861611854, 29.95312106759227]]]}
{‘ZipCode’: ‘70001’, ‘type’: ‘Polygon’,
‘coordinates’: [[[-90.2108570782496, 29.968073052931214],
[-90.2108570782496, 30.000181101129183],
[-90.1246895611469, 30.000181101129183],
[-90.1246895611469, 29.968073052931214],
[-90.2108570782496, 29.968073052931214]]]}
{‘ZipCode’: ‘70130’, ‘type’: ‘Polygon’,
‘coordinates’: [[[-90.0870713191487, 29.919017435693686],
[-90.0870713191487, 29.958422767573012],
[-90.05797153891734, 29.958422767573012],
[-90.05797153891734, 29.919017435693686],
[-90.0870713191487, 29.919017435693686]]]}
{‘ZipCode’: ‘70117’, ‘type’: ‘Polygon’,
‘coordinates’: [[[-90.05996712261795, 29.94874269738346],
[-90.05996712261795, 29.98618223039352],
[-90.02970573396384, 29.98618223039352],
[-90.02970573396384, 29.94874269738346],
[-90.05996712261795, 29.94874269738346]]]}
{‘ZipCode’: ‘70094’, ‘type’: ‘Polygon’,

‘coordinates’: [[[-90.27358393307792, 29.887534868954116],
[-90.27358393307792, 29.96983035798101],
[-90.12742976563132, 29.96983035798101],
[-90.12742976563132, 29.887534868954116],
[-90.27358393307792, 29.887534868954116]]]}
{‘ZipCode’: ‘70112’, ‘type’: ‘Polygon’,
‘coordinates’: [[[-90.0883222820655, 29.94478131481358],
[-90.0883222820655, 29.966285963050066],
[-90.06565602540627, 29.966285963050066],
[-90.06565602540627, 29.94478131481358],
[-90.0883222820655, 29.94478131481358]]]}