CURATING AI-READY DATASETS FOR EQUITY AND ENVIRONMENTAL JUSTICE: A DATA-CENTRIC AI CASE STUDY

Paridhi Parajuli[1], Rajat Shinde[1], Iksha Gurung[1], Manil Maskey[2], Rahul Ramachandran[2]

[1]The University of Alabama in Huntsville, AL, U.S.A
[2]NASA Marshall Space Flight Center, Huntsville, AL, U.S.A.

ABSTRACT

An equitable and environmentally just community is essential in order to avoid disproportionate burden borne by vulnerable communities. This need becomes pressing in the aftermath of an extreme event such as disaster or hazard when it is difficult for the governing bodies to implement resource allocation as per the need. Artificial Intelligence (AI) algorithms can help surface Equity and Environmental Justice (EEJ) issues when trained on EEJ datasets. However, curating AI-ready EEJ training datasets is challenging due to differences in factors such as heterogeneity, resolution, modality, and level of expertise in labeling. Additionally, EEJ issues involve sensitive information where uncertainties and errors could degrade the performance of AI algorithms. For eg. Error in seasonal crop yield information can highly affect the prediction of annual crop yield. To address these challenges, Data-centric AI (DCAI) methods are employed, which enhance AI algorithm performance even with limited training samples. DCAI prioritizes data quality, thereby reducing the adverse effects of uncertainties and errors during the model training process. This research proposes a novel dataset and benchmark for analyzing the effect of the Maui Wildfire of 2023 for Equity and Environmental Justice (EEJ) issues. The proposed dataset aligns with the concepts of DCAI such as annotation quality, data preprocessing, privacy, feature engineering, governance and provenance. We firmly believe that the proposed dataset would lay a foundation to implement robust and reliable modern AI algorithms for addressing EEJ issues.

Index Terms— Data-Centric AI, Equity and Environmental Justice (EEJ), Air Quality, Wildfire

1. INTRODUCTION

Equity and Environmental injustice refers to unequal distribution of environmental burdens as well as benefits to various communities based on their racial, social, economic and demographic characteristics. Addressing this inequality is crucial to promote an equal distribution of natural and man-made resources as per the needs of different communities. This is significantly evident while tackling extreme events such as natural disasters and hazards like wildfires. Such events cause relatively more burden on vulnerable communities, hence causing environmental injustice. However, the above-mentioned problems can be surfaced to a great extent by using AI approaches and rapid analysis at a big scale[1].

1.1. Data-centric AI

Data-centric AI (DCAI) approach focuses primarily on the data quality, quantity, and effective utilization of data for training an AI model using a few training samples. In DCAI, more emphasis is put on improving the quality of data as compared to developing AI model. With high-quality data and AI algorithms, there is an opportunity to surface the EEJ problems by designing novel strategies and implementation at scale to mitigate inequities in society. Geospatial Foundation Models (FMs)[1] can play a key role in achieving this target by using pre-trained models for developing new data products. Meanwhile, this is still far-fetched as a good quality dataset for training such models is lacking. To address this, data-centric AI techniques [2, 3] focusing on generating robust and reliable datasets can play a big role. Our research focuses on utilizing the data-centric AI tools for analyzing injustice from a spatio-temporal perspective. Additionally, we present validation of a proposed dataset for inference on the Segment Anything Model [4] by presenting preliminary benchmarking results for generating EEJ segmentation masks for various EEJ variables. Additionally, with data-centric algorithms such as [5] and workflows for responsible curation of datasets, accountability increases in the application of AI for policy designing and implementation. Readers are encouraged to follow some examples of recently published AI-ready data sets and data benches [6–8].

Our main contributions are:

- An AI-ready dataset based on data-centric concepts capable of training AI models for analyzing effect of wildfire with respect to EEJ issues.
- Benchmarking results for inferencing on SAM model using the proposed dataset for predicting EEJ segmentation masks.
- We present the effect of Maui wildfire on the socio-economic and physiological factors based on three case studies involving multiple EEJ variables.

[1] *https://www.earthdata.nasa.gov/dashboard/stories/hurricane-maria-and-ida*

2. CASE STUDY: MAUI WILDFIRE OF AUGUST 2023, HAWAII, THE U.S.

This research studies the environmental injustice caused by the devastating Maui wildfire of 2023 in the Hawaiian islands of the U.S. This 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 [2]. 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[3]. This catastrophe profoundly affected air quality, critical infrastructure, and economic activities[4], evident in a surge of unemployment claims from 130 to 2,705 cases per week[5]. Unveiling instances of environmental injustice exacerbated by the fire presents a unique challenge, necessitating a comprehensive understanding of the wildfire's far-reaching consequences. Our aim is to leverage AI to highlight instances of such environmental injustice, enabling the efficient and automatic identification of such issues in the future. The primary challenge in utilizing AI for addressing EEJ issues arises due to unavailability of good quality AI ready datasets for implementing state-of-the-art algorithms. The existing data is available in different spatio-temporal resolution, projection and coordinate reference systems attributed to the genesis from various sources. In this regard, remote sensing and satellite imagery derived proxies and indices play a crucial role to address this limitation.

2.1. Dataset Description

The proposed AI-ready dataset[6] comprises multiple channels corresponding to different attributes mentioned in Table 1. We have curated a diverse set of data including land cover, socio economic, air quality, demographic and topographical. Remote Sensing derived indices like Normalized Difference Vegetation Index (NDVI), Normalized Difference Water Index (NDWI) and satellite image based observation such as Aerosol Index, Nighttime light are collected on pre and post wildfire timestamps.

Data representing equity like demographics, topography, socioeconomic value are collected at only one timestamp owing to the limitation in frequency of Census data. In future, we plan to incorporate community curated datasets like OpenStreetMap or Humanitarian OSM maps based proxies for more frequent equity representation. The column "Frequency" indicates whether the data is collected temporally or not. For example, Racial category data is collected spatially only at one timestamp. On the other hand, remote sensing indices are derived temporally i.e for dates before and after the Maui wildfire. The pre datasets are selected from August 01 - August 07 whereas the post datasets are selected from August 18 - August 25, 2023.

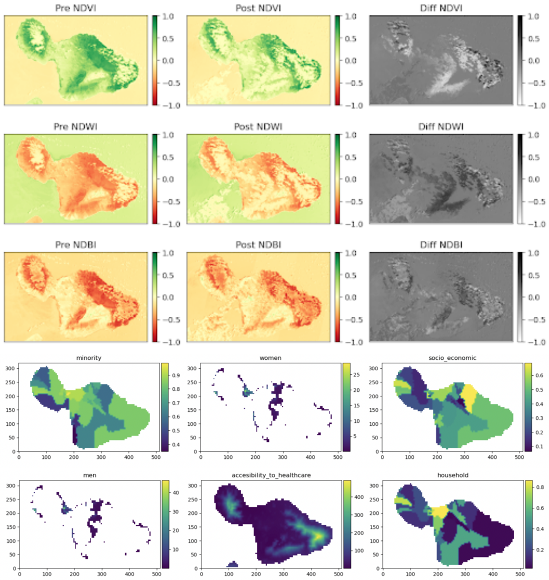


Fig. 1: Illustration of different channels of the Widfire AI-ready Dataset

Table 1: Description of different data sources with EEJ category and its attributes *(SEDAC: Socioeconomic Data and Applications Center; FIRMS: Fire Information for Resource Management System)*

Attribute	Frequency	Category(Source of the dataset)
White, Black, African American, American Indian and Alaska Native, Asian, Native Hawaiian, Pacific Islander, Hispanic	Constant	Racial (SEDAC)
Minority Population Household/Housing Characteristic Social Vulnerability Index	Constant	Socioeconomic Proxy (SEDAC)
Copernicus Global Digital Elevation Model 30m	Constant	Topography (Copernicus)
Women/Men/Age Under 5 Over 65 Elderly/Youth Population	Constant	Demographic Distribution (SEDAC)
Near-real time Aerosol Index	Pre/Post	Air Quality (Copernicus Sentinel-5P)
NDVI, NDWI, NDBI, NDMI	Pre/Post	Remote Sensing Indices
Night Time Light Data	Pre/Post	Proxy of Economy (VIIRS)
Fire Raster Mask	Pre/Post	Validation of Burn Scars (NASA FIRMS)

Figure 1 visualizes the temporal remote sensing derived indices along with EEJ categories. The proxies based on indices show a significant change in vegetation, water content, building and moisture over the study area. Also, it is evident that there is a diversity in equity as observed from illustration of racial, ethnicity, gender and socio-economic distribution over the study area. This justifies our strategy for a location

[2] *https://www.nytimes.com/interactive/2023/08/10/us/maui-wildfire-map-hawaii.html*

[3] *https://www.governing.com/work/lahaina-wildfire-victims-unemployment-claims-remain-uncertain*

[4] *https://dbedt.hawaii.gov/blog/23-47/*

[5] *https://www.mauinews.com/news/local-news/2023/11/mauis-unemployment-rate-shot-up-to-8-4-in-september*

[6] *We plan to host the data on a community contributed data platform such as Zenodo or HuggingFace Datasets in Croissant format.*

based equity and environmental injustice analysis with time.

2.2. Data-centric AI Concepts and Implementation

We applied following data-centric AI concepts:

- *Data Provenance*: The proposed dataset has been curated only using authoritative data sources in order to justify data provenance. Sources like Sentinel data products, US Census data, Social Vulnerability Index (SVI) data from Socioeconomic Data and Applications Center (SEDAC), and Fire Information for Resource Management System (FIRMS) are used.
- *Versioning*: We plan to extend this data and follow semantic versioning for further releases.
- *Data Quality*: Data quality checks have been performed on the proposed dataset to remove outliers and no data values.
- *Data Transformation*: Majority of the EEJ attributes such as racial and demographic are not as frequent as other remote sensing based datasets. In order to perform the analysis, proper data transformation is performed for such attributes.
- *Data Deployment*: In order to progress reproducible research and open science, the proposed dataset is planned to be hosted online so that other researchers can easily access these datasets in performing their research activities.

3. AI BENCHMARKING RESULTS

3.1. Validation Task Description

Figure 2 illustrates the curated AI-ready dataset and overview of benchmarking tasks. The task is to generate EEJ masks by mapping the channels from the AI-ready dataset to an EEJ variable for all the selected EEJ variables. The generated EEJ masks can be used for quantifying the effects of wildfire over a region with an EEJ lens.

3.2. Baseline Experiment

Our analysis is carried out by taking combinations of a difference variable and a racial/demographic variable. First, difference rasters for each temporal variable before and after the wildfire are calculated. The difference raster is stacked with a channel of spatially varying EEJ variable to investigate the effects of wildfire on a particular variable. Additionally, we calculate a fire mask for overlaying the regions affected by the wildfire. We stack up the fire mask channel with the two channels to form a data cube for a particular EEJ variable. Subsequently, all the channels are normalized in the range of $0 - 255$. This three channel data is saved as a RGB image.

We experiment the validity of inferencing using the pretrained Segment Anything model (SAM) by feeding the RGB image where R corresponds to *racial distribution*, G to *ndvi_diff* and B to *fire mask*. The SAM model generates an EEJ segmentation mask and a prediction Intersection over Union (IoU) score, highlighting the quantitative and qualitative relationship between the input three channels with respect to the EEJ variable.

3.3. Results

We performed comprehensive experiments for analyzing the EEJ based on multiple attributes, as described above. This section highlights the results in three subsections focusing on a particular EEJ problem and the insights generated based on the selected channels of the AI-ready dataset.

*3.3.1. **Case 1**: Spatio-temporal variation of Vegetation due to the wildfire with respect to the racial distribution*

Analysis: It is qualitatively evident that the NDVI values have reduced in the places with majority of the total populations (Figure 3a). The EEJ segmentation mask implies areas with decrease in values for different racial communities (Figure 3b). The validation is based on the alignment of the EEJ segmentation mask with the fire raster mask justifying that the reduction in NDVI is due to wildfire. Inference based on SAM achieves a prediction IoU of 0.8802.

*3.3.2. **Case 2**: Measuring change in socio-economic activity with the night-time light as proxy*

Analysis: Mostly negative night-light time difference in the regions of fire imply reduction in economic activity assuming night-time light as a proxy for industrial activities (Figure 4). This is justified with the EEJ segmentation mask and its overlap with actual fire location.

*3.3.3. **Case 3**: Analyzing infrastructural damage due to wildfire using change in NDBI with respect to total population*

Analysis: Mostly negative difference values of NDBI mask with total population is observed (Figure 4). The EJ classification mask validates this difference with the fire rasters implying the destruction is due to the wildfire. In this case, the prediction IoU achieved is 0.9003.

*3.3.4. **Discussion:***

There is around 77% of the area that observed reduction in the vegetation cover (based on the NDVI) with major reduction observed in places with high racial minority populations. Also, 15% areas observed reduction in Night-time light data whereas 50% areas observed reduction in NDBI is observed due to the Maui wildfire over the region of fire.

4. SUMMARY

The future work involves extending this work by adding more attributes to the dataset. Also, we foresee the addition of datasets related to more disasters and hazards to curate a novel data bench for progressing research in the field of machine learning for equity and environmental justice. The dataset will serve as a foundation for various machine learning tasks, enabling researchers to apply data-centric AI methods such as

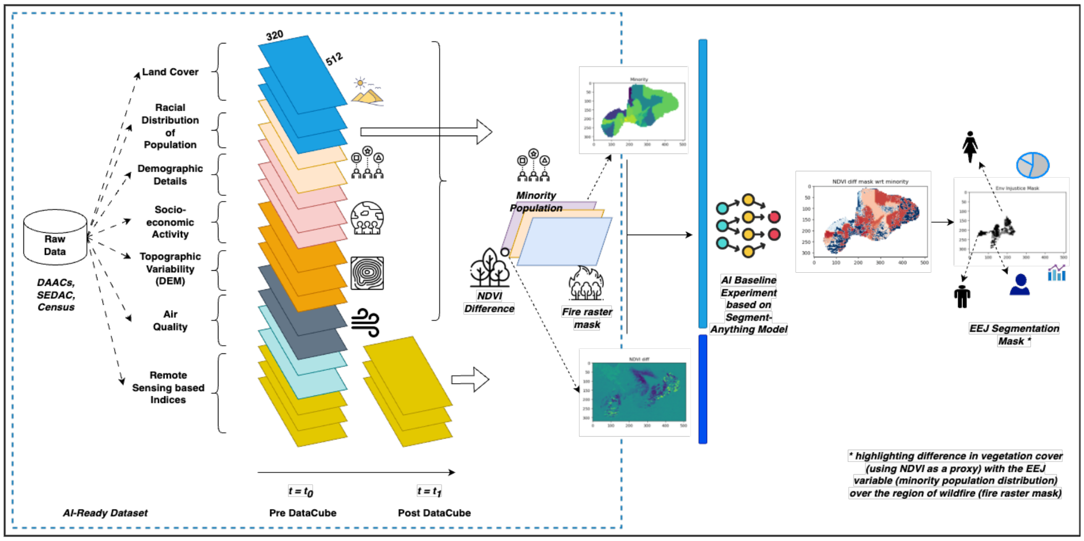
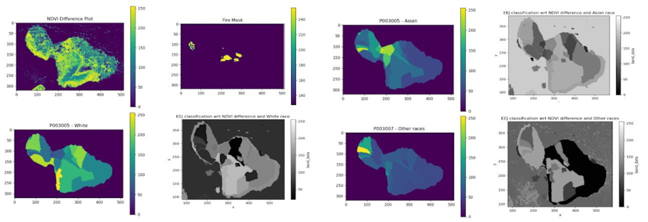

320
512
Land Cover
Racial Distribution of Population
Demographic Details
Socio-economic Activity
Topographic Variability (DEM)
Air Quality
Remote Sensing based Indices
Raw Data
DAACs, SEDAC, Census
Minority Population
NDVI Difference
Fire raster mask
Minority
NDVI diff
AI Baseline Experiment based on Segment-Anything Model
NDVI diff mask wrt minority
Env Injustice Mask
EEJ Segmentation Mask *
* highlighting difference in vegetation cover (using NDVI as a proxy) with the EEJ variable (minority population distribution) over the region of wildfire (fire raster mask)
$t = t_0$
$t = t_1$
AI-Ready Dataset
Pre DataCube
Post DataCube

Fig. 2: Illustration of the curated AI-ready dataset and benchmarking for analyzing the effects of an extreme event - Maui Wildfire in Hawaii, the U.S.


NDVI Difference Plot
Fire Mask
P003005 - Asian
EEJ classification wrt NDVI difference and Asian race
P003005 - White
EEJ classification wrt NDVI difference and White race
P003007 - Other races
EEJ classification wrt NDVI difference and Other races

(a) (b)

Fig. 3: Illustration of variation in NDVI due to wildfire w.r.t (a) "White community", (b) "Asian" and "Other races" (based on U.S. Census Data Variable P003002, P003005, and P003007)

data augmentation and confident learning. This dataset would open avenues for innovative developments of AI algorithms to surface EEJ issues for data-driven decision making.

5. REFERENCES

[1] 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.

[2] Manil Maskey, "Rethinking ai for science: An evolution from data-driven to data-centric framework," *Perspectives of Earth and Space Scientists*, vol. 4, no. 1, pp. e2023CN000222, 2023.

[3] 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.

[4] A Kirillov, E Mintun, N Ravi, H Mao, C Rolland, L Gustafson, T Xiao, S Whitehead, AC Berg, and WY Lo, "Segment anything. arxiv preprint arxiv: 230402643," 2023.

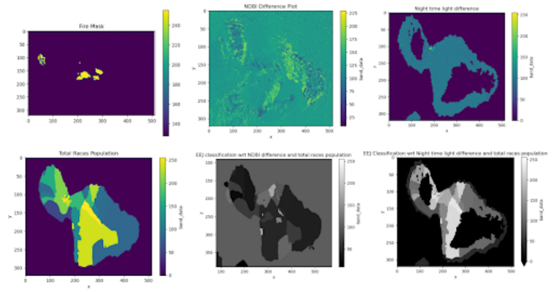

Fire Mask
NDBI Difference Plot
Night time light difference
Total Races Population
EEJ classification wrt NDBI difference and total races population
EEJ Classification wrt Night time light difference and total races population

Fig. 4: Illustration of effects on socio-economic activities and infrastructure due to wildfire (based on Night-time light difference and Normalized Difference Built-up Index (NDBI))

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

[6] Sungduk Yu, Walter M Hannah, Liran Peng, Mohamed Aziz Bhouri, Ritwik Gupta, Jerry Lin, Björn Lütjens, Justus C Will, Tom Beucler, Bryce E Harrop, et al., "Climsim: An open large-scale dataset for training high-resolution physics emulators in hybrid multi-scale climate simulators," *arXiv preprint arXiv:2306.08754*, 2023.

[7] Karthik Kashinath, Mayur Mudigonda, Sol Kim, Lukas Kapp-Schwoerer, Andre Graubner, Ege Karaismailoglu, Leo Von Kleist, Thorsten Kurth, Annette Greiner, Ankur Mahesh, et al., "Climatenet: An expert-labeled open dataset and deep learning architecture for enabling high-precision analyses of extreme weather," *Geoscientific Model Development*, vol. 14, no. 1, pp. 107–124, 2021.

[8] 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*, vol. 13, no. 6, pp. 3013–3033, 2021.