Environmental Science Basics III: Biodiversity, Agriculture; Land Use

This section provides an overview of key concepts related to Biodiversity and Ecosystems; Sustainable Agriculture; Land Use and Land Cover.

Ecology and Biodiversity

A key focus in environmental data analytics is the study of living organisms, from large mammals to plants, down to communities of microorganisms, along with their ecosystems, such as forests, grasslands, and aquatic environments, and the interactions within these systems. Environmental data analytics may also focus on **biodiversity**, which is technically defined as the variability among living organisms from all sources, including terrestrial, marine, and aquatic ecosystems, as well as the ecological complexes they are part of. This variability is critical because it supports ecosystem function, ensuring the continuous flow of resources like food, clean water, and oxygen, and provides genetic diversity that enhances species adaptation and resilience to environmental changes.

Biodiversity and Ecosystem-related Data

In Environmental Data Analytics, biodiversity and ecosystem-related issues are often addressed by leveraging **large datasets** to analyze and model interactions between species, ecosystems, and environmental factors. The data used in these analyses often falls into the following categories:

1. Environmental Data:

- **Climate and Weather Variables:** Includes data on temperature, precipitation, wind speed, and solar radiation, which influence species distributions and ecosystem dynamics.
- **Hydrological Data:** Information on the water cycle, such as river flows, groundwater levels, and water quality parameters like pH and nutrient concentrations.
- **Air, Water, and Soil Quality:** Pollution metrics, including CO_2 levels, particulate matter, and chemical pollutants in air, water, and soil.

2. Single Species Occurrence Data, such as:

- **Species Observation Records:** Point-in-time data detailing where and when a specific species was observed, typically collected via field surveys, citizen science platforms, or biodiversity databases.

3. Single Species Movement Data, such as:

- **Tracking and Migration Patterns:** Data obtained from GPS collars, satellite tracking, or tagging that captures the movement and migration of a species over time and space.
- **Home Range and Territory Use:** Spatial data on the area utilized by an individual or population of a species within a given time frame.

4. Population-Level Data, such as:

- **Population Density and Distribution:** Data on the number of individuals of a species in a given area, and how the population is spread across regions. A common method of data collection is ***transect surveys*** where researchers walk or fly along fixed paths (transects) and record the number of individuals encountered, allowing for estimation of density and distribution in a specific area.
- **Demographic Data:** Information on population dynamics, including birth rates, death rates, and age structure for specific species populations. **Camera traps** are commonly used for this purpose.

A camera trap.

5. Community-Level Data, such as:

- **Species Richness and Diversity Metrics:** Measures of species diversity within an ecosystem, such as species richness and abundance. A common method of

data collection is **Quadrat Sampling** where researchers mark out small, square plots (quadrats) in the study area and count how many different species (and how many individuals of each species) are found within the quadrat. This is then extrapolated to estimate species richness and abundance across the entire ecosystem.

- **Ecological Interaction Networks:** Data on interactions among species, such as predator-prey relationships, competition, and mutualism, which help in understanding community structure and ecosystem functioning. To gather such data, ecologists often watch species in the field to document interactions.

A vulture with electronic tag.

Single Species Occurrence Data, Single Species Movement Data, Population-Level Data, and Community-Level Data are typically collected through extensive fieldwork by ecologists, either through direct observation or with the help of tools such as camera traps, tracking devices like animal tags, or sampling from feces. Environmental data analytics specialists are usually not involved in the collection of such data, as it requires specialized ecological knowledge and fieldwork techniques. Instead, their primary role is to analyze the data to answer specific research questions or to generate insights.

Definition of key terms used in the above classification:

- **Species:** A group of organisms capable of interbreeding and producing fertile offspring, sharing common characteristics and classified under the same biological category.

- **Population-Level:** Refers to the study of a group of individuals of the same species living in a specific geographic area, focusing on their size, density, distribution, and dynamics over time.
- **Community-Level:** Refers to the analysis of multiple interacting species living within a shared environment, focusing on relationships like predation, competition, and symbiosis within an ecosystem.

Examples of Publicly Available Ecology and Biodiversity Datasets

Here are three examples of publicly available ecology and biodiversity datasets:

1. **Global Biodiversity Information Facility (GBIF):**
    - **Description:** A global database providing access to occurrence records of species collected from field observations, museum specimens, and citizen science platforms. It includes over 2 billion species occurrence data points from around the world.
    - **Data Type:** Species occurrence data, including location, date, and species names.
    - **Access:** GBIF
2. **LTER (Long Term Ecological Research Network) Data Portal:**
    - **Description:** A collection of long-term datasets from ecological research sites across diverse ecosystems, including forests, wetlands, and deserts. The data include measurements on climate, species abundance, nutrient cycling, and ecosystem changes over time.
    - **Data Type:** Long-term ecological data on ecosystems, climate, and biodiversity.
    - **Access:** LTER Data Portal
3. **eBird Database:**
    - **Description:** A massive citizen science project where birdwatchers from around the world submit sightings. eBird provides data on bird species occurrence, abundance, and migration patterns, and is frequently used for studying bird biodiversity and distribution.
    - **Data Type:** Species occurrence, bird migration patterns, and abundance data.
    - **Access:** eBird

Types of Questions Addressed by EDA related to Ecology and Biodiversity

Here are examples of typical questions in ecology and biodiversity that are commonly addressed using Environmental Data Analytics (EDA):

1) ***Question*:** How will climate change impact the distribution of a specific species in the next 50 years?

   ***Common Approach*:** EDA uses species data from distribution models (SDMs) combined with climate projection data to predict future shifts in species ranges based on environmental factors like temperature and precipitation.

2) ***Question*:** What factors are driving the decline of an endangered species, and what conservation measures can reverse this trend?

   ***Common Approach*:** EDA analyzes time-series data on population size, habitat quality, and human impacts (e.g., land use, hunting) to identify correlations and drivers of population decline, informing targeted conservation strategies.

3) ***Question*:** Where are biodiversity hotspots located, and how are they affected by habitat fragmentation and human activities?

   ***Common Approach*:** EDA performs spatial analysis on biodiversity richness metrics and overlays land-use data to identify fragmentation patterns and human impacts on critical biodiversity hotspots.

4) ***Question*:** How do predator-prey relationships within an ecosystem change following the introduction of an invasive species?

   ***Common Approach*:** EDA constructs and analyzes ecological interaction networks using data from field observations and tracking studies to assess changes in predator-prey dynamics due to the presence of invasive species.

5) ***Question*:** What is the impact of agricultural pollution on the biodiversity of nearby freshwater ecosystems?

   ***Common Approach*:** EDA applies spatial analysis and statistical modeling on water quality data and biodiversity metrics to determine the correlation between pollution levels and changes in species diversity and abundance in freshwater ecosystems.

Sustainable Agriculture

Sustainable Agriculture refers to farming practices that meet current food and textile needs without compromising the ability of future generations to meet their own needs. Sustainable agriculture practices aim to minimize environmental impact by promoting biodiversity, improving soil health, conserving water, and reducing the use of chemical inputs like pesticides and fertilizers.

Precision agriculture is closely related to sustainable agriculture as it enables farmers to optimize the use of resources—such as water, fertilizers, and pesticides—by tailoring management practices to specific field conditions, thereby reducing waste, minimizing environmental impact, and enhancing long-term productivity and sustainability.

Precision agriculture is a data-driven farming approach that uses advanced technologies such as GPS, remote sensing, and IoT sensors to monitor and manage variability in crops, soil, and environmental conditions, optimizing resource use and maximizing productivity at a granular, site-specific level.

An lot sensor used in a greenhouse for precision agriculture.

Nature of Data in Sustainable Agriculture

In sustainable agriculture, Environmental Data Analytics involves processing and analyzing various types of data to enhance farming practices and minimize environmental impact. Examples of data related to EDA in sustainable agriculture include:

1. **Crop Data**
    - **Crop Health Index (NDVI) (Static Image, or Image Time Series)**

- **Format**: Raster images (multispectral or hyperspectral) showing NDVI over time.
- **Measurement**: Derived from remote sensing imagery (e.g., satellites, drones).
- **Usage**: Monitors crop vigor, biomass, and photosynthetic activity, enabling farmers to optimize growth and detect early signs of stress in crops.

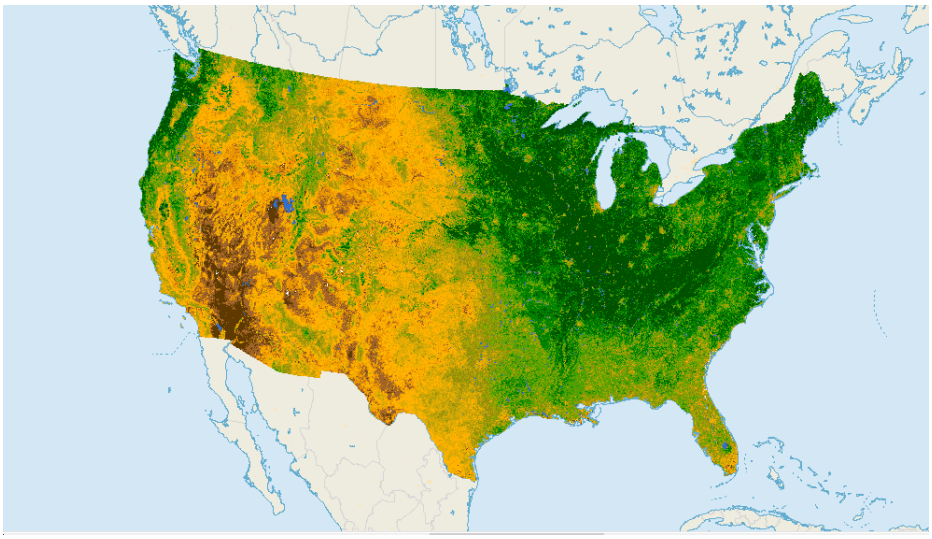
- **Crop Yield (Time Series, or Spatial Data)**
  - **Format**: Time series or spatial data representing yield in tons per hectare.
  - **Measurement**: Estimated from harvest data or calculated using remote sensing techniques such as satellite or drone imagery.
  - **Usage**: Provides insight into productivity and helps evaluate the effectiveness of farming practices.

2. **Soil Data**
   - **Soil Moisture (Time Series, Map)**
     - **Format**: Time series data or spatial maps (raster) that depict soil moisture content across time and space.
     - **Measurement**: Measured using sensors like tensiometers at various depths, or derived from remote sensing technologies (e.g., satellite-based Synthetic Aperture Radar, SAR).
     - **Usage**: Optimizes irrigation strategies and monitors water stress levels in crops, ensuring water efficiency.
3. **Land Use and Land Cover Data (Image, Map)**
   - **Format**: Geospatial raster images or vector maps showing land use classifications (e.g., cropland, forest, urban areas).
   - **Measurement**: Derived from satellite or drone imagery and classified into land use types using algorithms such as Random Forest or Support Vector Machines (SVM).
   - **Usage**: Assesses land cover changes, monitors deforestation, and evaluates the impact of land use on ecosystems and sustainability.
4. **Water Usage Data (Time Series)**
   - **Format**: Time series data detailing water extraction rates and usage volumes.

- **Measurement**: Collected via flow meters in irrigation systems or estimated through remote sensing models like evapotranspiration models.
- **Usage**: Helps optimize irrigation schedules and ensures water is used efficiently, reducing waste and conserving resources.

Drones are highly effective data collection tools in precision agriculture because they can rapidly cover entire fields and provide high-resolution, on-demand data for real-time monitoring and decision-making.

Tensiometer for soil moisture measurement at a specific location and depth.



NDVI (Normalized Difference Vegetation Index) map of the United States.

Examples of Publicly Available Sustainable Agriculture Datasets

Here are two popular examples of publicly available datasets for Environmental Data Analytics (EDA) in sustainable agriculture:

1) **MODIS (Moderate Resolution Imaging Spectroradiometer) NDVI Data**
   - **Description**: The MODIS NDVI dataset provides global vegetation index data derived from satellite imagery. It offers high temporal resolution data (every 1-2 days) and moderate spatial resolution (250m-1km), making it ideal for monitoring crop health, vegetation cover, and land use changes.
   - **Format**: Available as raster images in GeoTIFF format, often used in time series for vegetation analysis.
   - **Source**: NASA Earth Observing System Data and Information System (EOSDIS) MODIS NDVI Dataset
   - **Usage**: Widely used to assess crop health, monitor drought impacts, and study land cover changes.
2) **SoilGrids by ISRIC (International Soil Reference and Information Centre)**
   - **Description**: SoilGrids is a global soil information system providing predictions for soil properties like pH, organic carbon, and moisture content at different depths (0-2m). The data is presented at a spatial resolution of 250 meters, which allows for detailed soil quality assessment at the regional and global scale.

- **Format**: Available as raster layers (GeoTIFF) and CSV format for point-based data.
- **Source**: ISRIC - World Soil Information SoilGrids Dataset
- **Usage**: Used in sustainable agriculture to assess soil quality, plan crop rotations, and manage fertilization and irrigation practices based on soil conditions.

Types of Questions Addressed by EDA in Sustainable/Precision Agriculture

Here are examples of typical questions in sustainable agriculture that can be addressed through Environmental Data Analytics (EDA):

1. **Crop Health Monitoring**
   - *Question*: How can I detect early signs of crop stress due to drought or pests using satellite data?
   - *Common Approach*: EDA processes NDVI or other vegetation indices derived from satellite imagery to identify areas with poor crop health or stress.
2. **Water Management Optimization**
   - *Question*: How can I optimize irrigation schedules to maximize crop yield while minimizing water use?
   - *Common Approach*: EDA uses soil moisture sensors, weather data, and evapotranspiration data to develop efficient irrigation plans that reduce water wastage.
3. **Soil Quality Assessment**
   - *Question*: What is the spatial variation of soil nutrients across my farm, and how can I adjust fertilization strategies accordingly?
   - *Common Approach*: EDA analyzes soil test results or remote sensing data to map nutrient levels and inform "precision agriculture" strategies for fertilizer application.
4. **Climate Resilience**
   - *Question*: How will future climate scenarios affect the productivity of specific crops in my region over the next decade?
   - *Common Approach*: EDA integrates historical climate data and data from predictive models to evaluate how changing temperatures, rainfall patterns, and extreme weather events could affect crop yields.

5. **Sustainability and Carbon Footprint**
    - *Question*: How can I measure and reduce the carbon footprint of my farming practices while maintaining productivity?
    - *Common Approach*: EDA tracks greenhouse gas emissions (e.g., from soil and livestock) and carbon sequestration through soil and biomass data to suggest practices that lower emissions while sustaining yields.

Land Use and Land Cover

Land Use refers to the human activities or economic functions associated with a particular piece of land, such as agriculture, urban development, etc.

Land Cover refers to the physical material on the Earth's surface, such as vegetation, water bodies, or artificial structures like roads and buildings.

Note that some use these two terms interchangeably.

For a small, easily accessible piece of land, determining land use and land cover may not be difficult, as one can simply visit the area to observe and document its characteristics. However, this approach is not scalable to larger regions, especially those that include remote or inaccessible areas. Mapping **Land Use** and **Land Cover** on large scales is challenging due to the Earth's vast and complex surface, as well as the dynamic nature of land-use patterns that are constantly changing due to human activities, natural events, and environmental shifts over time.

Satellite imagery and remote sensing data provide vast amounts of information about **Land Use** and **Land Cover**, but processing these datasets to accurately distinguish between different types of land cover and land use requires advanced algorithms, machine learning models, and spatial data analysis. Environmental data analytics becomes essential in handling the scale, variability, and multi-source nature of the data, enabling automated classification, pattern detection, and continuous monitoring across large regions.

In **Land Use** or **Land Cover** identification within Environmental Data Analytics, the common problem is classifying or categorizing different types of land use or cover over large geographical areas.

- **Input:** The input typically consists of raw satellite imagery, remote sensing data, or aerial photographs. This data often contains pixel-level information and temporal data.
- **Output:** The output is usually a classified map or dataset where each pixel or region is labeled with specific land use or land cover categories (e.g., forest, water body, urban area, agriculture). These outputs help in understanding

spatial patterns, tracking changes over time, and supporting further environmental or ecological analysis.

Land Use and **Land Cover** identification is typically a **supervised learning task**, though unsupervised methods can also be applied in some cases. In supervised learning, the algorithm is trained on labeled data, where each pixel or segment of the input data (e.g., satellite imagery) is associated with a known land use or land cover category.

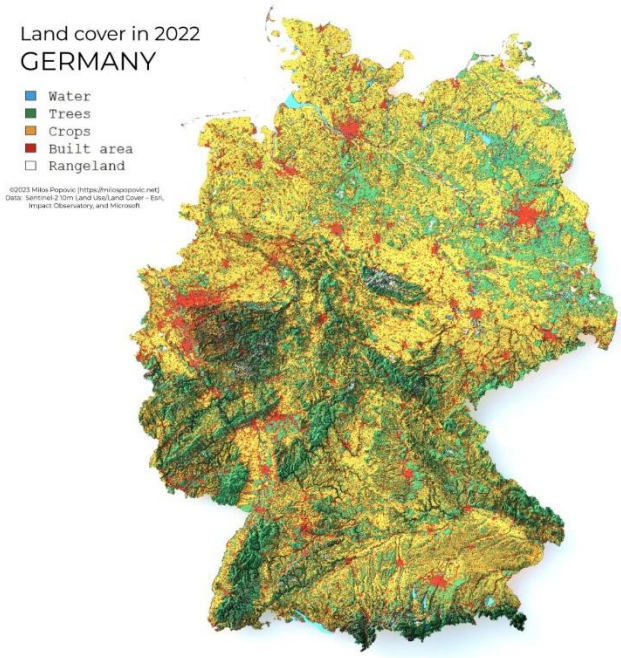
In the supervised case, **labeled data** is usually generated through a combination of methods:

1. **Ground Truthing**: Field surveys and observations are conducted to manually classify specific areas, though this is often limited to accessible regions.
2. **Expert Interpretation**: Experts manually interpret high-resolution satellite images or aerial photographs to label land cover types. This can involve visual inspection and marking of known categories, such as forests, urban areas, or agricultural fields.
3. **Existing Maps and Datasets**: Previously validated land use or land cover maps can be used as labeled datasets for training models on similar regions or for time-series analysis.
4. **Crowdsourcing**: In some cases, platforms like Google Earth Engine enable public participation in labeling imagery, which can contribute to generating large, labeled datasets.

In the context of Environmental Data Analytics, apart from the identification of Land Use or Land Cover, several **other machine learning tasks** are commonly performed on similar input datasets:

1. **Change Detection**: This involves analyzing satellite imagery or remote sensing data over different time periods to detect and quantify changes in land use or land cover (e.g., deforestation, urban expansion). Supervised or unsupervised learning models are used to identify shifts in land cover classes over time.
2. **Prediction and Forecasting**: Predictive models can be built to forecast future land use changes based on historical data and influencing factors such as population growth, economic development, or environmental policies. Regression and time-series forecasting techniques are often employed in these tasks.
3. **Anomaly Detection**: Anomalies or outliers in land use patterns, such as illegal logging or unexpected changes in natural habitats, can be detected using ML techniques like clustering, outlier detection, or autoencoders.

4. **Land Use Optimization**: ML models can help optimize land use patterns for environmental, economic, or social benefits, such as determining the best locations for conservation efforts, agriculture, or urban development based on multiple factors.



Land Cover map of Germany obtained from analyzing satellite data.

Examples of Publicly Available Datasets that can be Used for Land Use/Cover Studies

There are several publicly available datasets that are commonly used as input to **Land Use** and **Land Cover (LULC)** machine learning studies. These datasets typically include satellite imagery, remote sensing data, and labeled land cover maps. Here are some widely used ones:

1. Landsat Data (USGS)

- **Source**: United States Geological Survey (USGS)
- **Description**: Landsat satellites provide a long-term historical record of Earth's surface, with imagery available from 1972 onwards. It is frequently used for land cover classification, change detection, and environmental monitoring.
- **Access:** https://earthexplorer.usgs.gov/

- **Use Cases**: Land use classification, forest cover monitoring, urbanization studies.

2. Sentinel-2 (ESA)

- **Source**: European Space Agency (ESA)
- **Description**: Sentinel-2 provides high-resolution (up to 10m) multispectral imagery and is part of the Copernicus program. The data is used extensively for vegetation, soil, and water cover monitoring.
- **Access**: https://scihub.copernicus.eu/
- **Use Cases**: Agricultural monitoring, land cover change detection, urban land use mapping.

3. MODIS (Moderate Resolution Imaging Spectroradiometer)

- **Source**: NASA
- **Description**: MODIS provides global coverage at moderate spatial resolution (250m to 1km) and is useful for large-scale environmental studies, including vegetation cover, land use, and climate change.
- **Access**: NASA's Earthdata
- **Use Cases**: Global land cover change, environmental monitoring, forest and vegetation health.

4. Global Land Cover (GLC) Dataset

- **Source**: National Geomatics Center of China (NGCC)
- **Description**: GLC is a high-resolution (30m) global land cover dataset that includes multiple land cover classes such as forest, cropland, grassland, and built-up areas.
- **Access**: GlobeLand30
- **Use Cases**: Global land cover classification, deforestation studies, biodiversity assessment.

5. Global Forest Change Dataset

- **Source**: University of Maryland and Google
- **Description**: This dataset provides information on global forest cover change, including forest loss, gain, and other related metrics from 2000 onwards, derived from Landsat imagery.
- **Access**: https://earthenginepartners.appspot.com/science-2013-global-forest

- **Use Cases**: Deforestation monitoring, forest cover change analysis, carbon sequestration studies.

6. OpenStreetMap (OSM)

- **Source**: OpenStreetMap
- **Description**: OSM is a community-driven map of the world, containing detailed vector data on human infrastructure, such as roads, buildings, and land use. It is often used as complementary data in land use studies.
- **Access**: OpenStreetMap
- **Use Cases**: Urban land use classification, transportation network analysis, infrastructure mapping.

Disaster Management

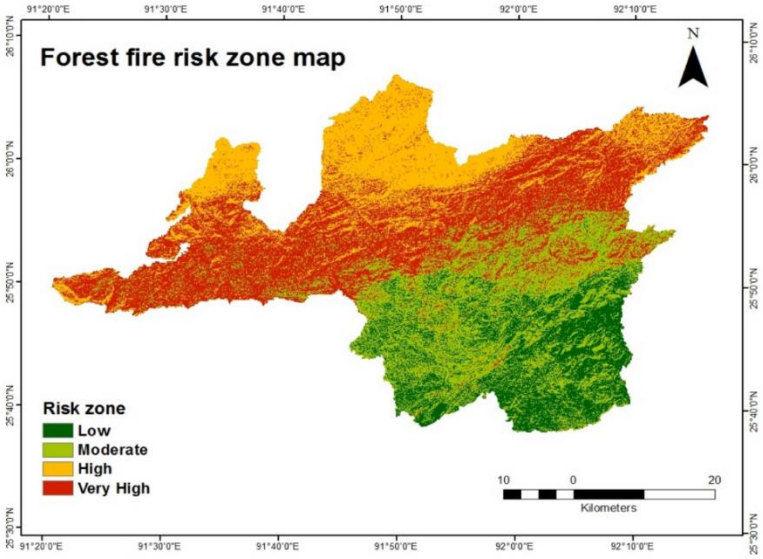
Environmental Data Analytics can play a significant role in **disaster management** by enabling better **preparedness**, and **response** efforts through the use of real-time data, advanced analytics, and machine learning. Here are several ways Environmental Data Analytics can assist in disaster management:

Risk Assessment and Mapping: By integrating geospatial and environmental data, EDA can identify areas most vulnerable to disasters and create risk maps. These maps help urban planners and decision-makers implement mitigation strategies, such as building flood defenses or reinforcing infrastructure in high-risk areas

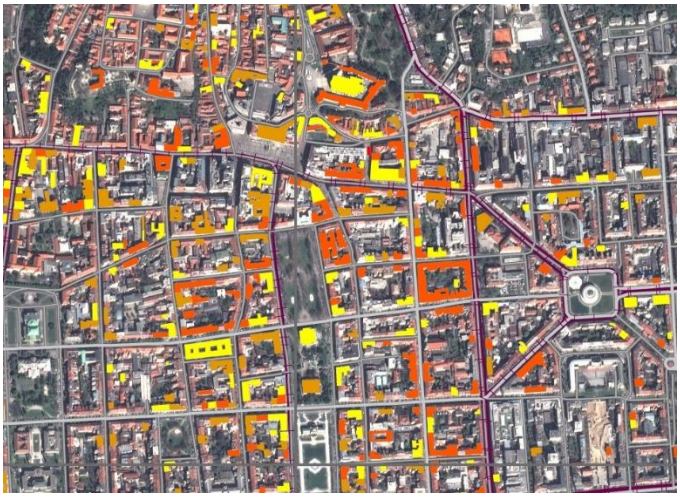
Early Warning Systems: By analyzing environmental data from satellites, sensors, and historical records, EDA can help develop predictive models for natural disasters such as hurricanes, floods, and wildfires. These models can provide early warnings, allowing authorities and communities to take preventive measures or evacuate in time.

Real-Time Monitoring: EDA can process and analyze data from various sources (e.g., weather stations, IoT sensors, remote sensing) in real time, providing situational awareness during disasters. For example, flood levels or fire spread can be monitored, enabling quicker response to mitigate impacts.

Damage Assessment: Post-disaster, EDA can analyze satellite imagery, drone data, and ground reports to assess the extent of damage in affected areas. This helps governments and organizations prioritize recovery efforts, restore services, and allocate funds effectively.



Forest fire risk mapping using earth observation datasets and machine learning n the mountainous terrain of Northeast India.



Damage assessment of buildings in Zagreb, Croatia, following the earthquake, using aerial imagery and machine learning.

Examples of Publicly Available Datasets that can be Used for Disaster Management Studies

1. **xBD: A Dataset for Assessing Building Damage from Satellite Imagery**
   - **Source**: Maxar and AWS

- **Description**: xBD is a large-scale dataset of satellite imagery containing annotations for building damage caused by natural disasters such as hurricanes, wildfires, and earthquakes. The dataset provides pre- and post-disaster images, along with labeled damage levels, making it ideal for machine learning tasks like damage detection, segmentation, and classification.
- **Access**: https://xview2.org/
- **Use Cases**: Building damage assessment, post-disaster recovery, and automated damage classification.

2. **Hurricane Harvey Flood Extent Dataset**
   - **Source**: NASA Earth Science Data Systems
   - **Description**: This dataset contains satellite-derived flood extent maps from Hurricane Harvey, focusing on Texas, USA. It includes imagery and floodwater data based on synthetic aperture radar (SAR) observations, which can be used to train machine learning models for flood detection and water body segmentation.
   - **Access**: https://data.nasa.gov/dataset/Hurricane-Harvey-Flood-Extent/
   - **Use Cases**: Flood detection, water body segmentation, and flood risk prediction.