Uncovering Optimal Solar Site Locations using Autoencoder and Clustering with Application in India

OLLSCOIL NA GAILLIMHE
UNIVERSITY OF GALWAY

Smitesh Nitin Patil
School of Computer Science
University of Galway

Supervisor(s)
Dr.Karl Mason

In partial fulfillment of the requirements for the degree of
MSc in Computer Science (Data Analytics)
[[Date of submission]]

DECLARATION I, Smitesh Nitin Patil, hereby declare that this thesis, titled "Uncovering Optimal Solar Site Locations using Autoencoder and Clustering with applications in India", and the work presented in it are entirely my own except where explicitly stated otherwise in the text, and that this work has not been previously submitted, in part or whole, to any university or institution for any degree, diploma, or other qualification.

Signature: _____

Abstract

This project delves into the integration of Autoencoders and clustering techniques within the framework of GIS (Geographical Information System) data to pinpoint optimal locations for Solar PV (Photo-Voltaic) installations. By harnessing advanced machine learning methodologies in tandem with spatial analysis, this research aims to carve out a novel approach, distinct from studies previously undertaken in this field, to the best of the authors' knowledge. Through the analysis of diverse environmental, climatic, and topographical factors, the proposed methodology furnishes a holistic solution for discerning areas with peak solar energy potential. The outcomes not only underscore the efficacy and robustness of the suggested approach but also highlight its prospective applications in the wider scope of renewable energy planning.

Keywords: Geospatial Information, Unsupervised Learning, Self-supervised Learning, Analytical-Hierarchical Process, Renewable energy, Site Selection, Spatial Analysis, Sustainability.

Contents

List of Figures

List of Tables

Acronyms

MLP Multi-Layer Perceptron

MCDMs Multi-Criteria Decision-Making Methods

AHP Analytical Hierarchy Process

NREL National Renewable Energy Laboratory

SRTM Shuttle Radar Topography Mission

DEM Digital Elevation Model

USGS United States Geological Survey

OSM OpenStreeMap

DNI Direct Normal Irradiance

GHI Global Horizontal Irradiance

DHI Direct Horizontal Irradiance

NSRDB National Solar Irradiance Database

CR Consistency Ratio

CI Consistency Index

RI Random Consistency Index

SOM Self-Organizing Map

BMU Best Matching Unit

PV Photo-Voltaic

NASA National Aeronautical and Space Agency

GIS Global Information System

PCA Principle Component Analysis

RBMs Restricted Boltzmann machines

CNN Convolutional Neural Network

CNNs Convolutional Neural Networks

Chapter 1

Introduction

1.1 Motivation

The global transition from fossil fuel-based energy sources to sustainable alternatives like wind and solar is paramount for the 21st century. The incentives for deploying these renewable sources are considerable. These resources are natural, free, abundant, and replenishable. Solar energy, generated from photovoltaic cells, requires consistent high solar irradiance throughout the year to be profitable. Tropical regions, like parts of India, benefit from abundant sunlight year-round.

India's energy demands are escalating. It stands as the third-largest producer of electricity globally, following the United States and China[5]. Presently, India's energy sector leans heavily on fossil fuels, with sources like coal fulfilling three-quarters of the country's energy needs. Nevertheless, India is making significant investments in solar and hydropower projects. The nation's committed to ensuring renewable energy sources account for 50% of energy consumption by 2030 and aspires to achieve net zero by 2070, as declared during the COP26 summit in 2021[6]. This commitment is evident as, between 2017 and 2021, India's solar power production capacity tripled, placing it third in global solar capacity

rankings[7].

Given the task's significance, it's crucial to rapidly identify new locations for renewable energy generation plants. The Indian government's national energy policy prioritizes solar and hydroelectric power generation. Situated between latitudes 20.5937° N and 78.9629° E, India's temperate and tropical climate conditions ensure high solar irradiance levels.

Before pinpointing promising regions for solar farms, several factors require careful consideration: the slope gradient of the terrain, proximity to urban centers, and the presence of conservation areas. Historically, scientific studies focusing on solar Photo-Voltaic (PV) plant installations, which leverage GIS data, have leaned towards the use of Multi-Criteria Decision-Making Methods (MCDMs) to evaluate these factors[7, 1, 8, 9, 2]. These studies predominantly employed the Analytical Hierarchy Process (AHP) as their chosen MCDMs technique to determine the relevant criteria. This research, however, ventures into exploring novel unsupervised learning methods. These methods draw parallels with techniques employed by researchers like Chang et al., who used them for monitoring landslide susceptibility using geospatial data[10]. Specifically, this study emphasizes the use of Autoencoders and clustering techniques to pinpoint regions that hold importance for the establishment of solar PV plants.

Although similar research has been conducted in India, many of these studies faced constraints due to the limited resolution of spatial data[11, 2, 12]. They also primarily relied on MCDMs for classification. The data underpinning this study is sourced from a diverse array of institutions, including the National Renewable Energy Laboratory (NREL) for solar irradiance, the Digital Elevation Model (DEM) provided by the Shuttle Radar Topography Mission (SRTM) spearheaded by the United States Geological Survey (USGS), and OpenStreeMap (OSM). The latter offers detailed insights into land use, protected reserves, water bodies, urban

centers, and transportation networks.

1.2 Purpose

The primary objectives of this study are:

1. To develop an innovative approach utilizing unsupervised and self-supervised learning methodologies for pinpointing optimal geolocations for the establishment of solar PV plants.

2. While prior studies on solar PV site selection in India were conducted with a limited scope, often relying on data with low spatial resolution (greater than 1000 meters), this research aims to leverage datasets with superior spatial resolution (ranging from less than 10 meters to 30 meters).

3. To the student's best knowledge, no previous studies have employed self-supervised Autoencoder and clustering based classification for PV sites on such a comprehensive scale.

1.3 Research Questions

1. Given the vast and varied sources of data (e.g., NREL, DEM from SRTM, OSM), how can they be effectively integrated to yield the most comprehensive insights for solar site selection?

2. Is it feasible to employ a novel technique that combines autoencoders and clustering to identify optimal areas for solar PV plant installations, based on data derived from the aforementioned sources?

Chapter 2

Background and Related Work

This chapter highlights the methods used by researchers previous in the domain in form of a literature review.

2.1 Criteria and Factors Affecting Decision-making

Selecting the right features for predictive models is a pivotal step in making informed decisions about the feasibility of specific geolocations for solar PV plants. There have been numerous studies undertaken by researchers to determine the critical factors for classifying PV solar plant sites.

Colak et al. conducted a study to identify suitable locations for establishing photovoltaic power plants in the Malatya province of Turkey[1]. The authors employed 11 layers of GIS data to pinpoint the most favorable sites. These layers encapsulated various factors that influence the decision-making process, such as:

1. **Solar Energy Potential:** Gauging the solar potential of a region is paramount. This metric essentially dictates the energy yield of a region when equipped with a photovoltaic power plant.

2. **Slope:** The terrain's slope plays a crucial role in the decision-making process. A more level terrain is preferred for the installation of PV panels, ensuring optimal exposure and ease of maintenance.

3. **Transformer Centers and Energy Transmission:** Transmitting electricity over vast distances without the appropriate energy infrastructure results in significant energy losses. Hence, having a power transmission system in place is crucial when considering a location for a new plant.

4. **Land Cover:** Certain lands designated as nature reserves, tribal habitats, or for other specific purposes are legally off-limits for energy generation activities. It's imperative to factor in these designations when choosing a site.

5. **Residential Areas:** Constructing a solar plant near an urban center might pose future challenges, especially with the continuous expansion of urban sprawl. Conversely, having a PV solar plant in proximity to urban centers can mitigate transmission losses. This duality necessitates a balanced consideration.

For data preprocessing, various hardset conditions were set to restrict certain areas like slope elevation of land cannot be more than 20 percent, distance to road, rail network should be more than 0.1 km, no residential areas nearby and proximity to energy transmission network.

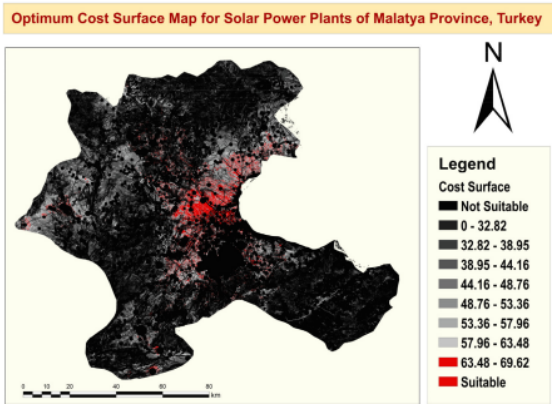


Figure 2.1: Suitable sites for Solar Powerplant wiht cost factor by author Colak et al.[1]

A similar study was conducted by Al Garni et al. in Saudi Arabia[8]. The available land was categorized into five classes: least suitable, marginally suitable, moderately suitable, highly suitable, and most suitable. The decision-making process for site selection unfolded in three phases:

1. Setting decision criteria and restrictions for the site selection study.
2. Prioritizing sites with high solar potential.
3. Conducting an analysis on the prioritized regions for informed decision-making.

Like the aforementioned study, the authors relied on GIS data provided by NREL, selecting attributes that determined the criteria for site selection. These included DEM, Solar irradiation, and Air Temperature. Broadly, these factors can be divided into two categories: technical (factors affecting energy production) and economical (factors influencing the economic viability of the project).

Zoghi et al. proposed dividing the factors into four major categories for their case study carried out in Isfahan province, Iran[9]:

1. **Environmental:** Land use, Protected Areas, Wetlands, and Water Resource.

2. **Geomorphological:** Elevation, Slope, and Aspect.

3. **Location:** Distance to City, Distance to Power line, and Distance to Transport network.

4. **Climatic:** Sunshine, Cloudy Days, Dusty Days, Solar Radiation, Rainy and Snowy Days, and Humidity.

The study carried out by Saraswat et al. (2021) represents the most elaborate case study for site selection of solar PV plants in India, to the best of the student's knowledge[2]. A significant limitation of this study is the data's spatial resolution. With a resolution of around 1000m, it is not well-suited for detailed DEM modeling and other attributes. Consequently, the solar farm suitability map produced in this study lacks granularity at a spatial level. However, various databases, provided by USGS and NREL, offer spatial resolutions of 30m and can be employed to yield more accurate predictions.

Data for the study were sourced from multiple governmental bodies: NREL for solar radiation, DIVA-GIS for roads and inland water bodies, and the DEM model was provided by the United States Geological Survey (USGS). The factors were segmented into three categories: technical, socio-environmental, and economic.

1. **Technical:** Solar Radiation, Slope, Aspect, and Elevation.

2. **Socio-Environmental:** Distance from coastline, Distance from water bodies, airports, and Land use.

3. **Economic:** Distance from urban areas, roads, transmission lines, and power plants.

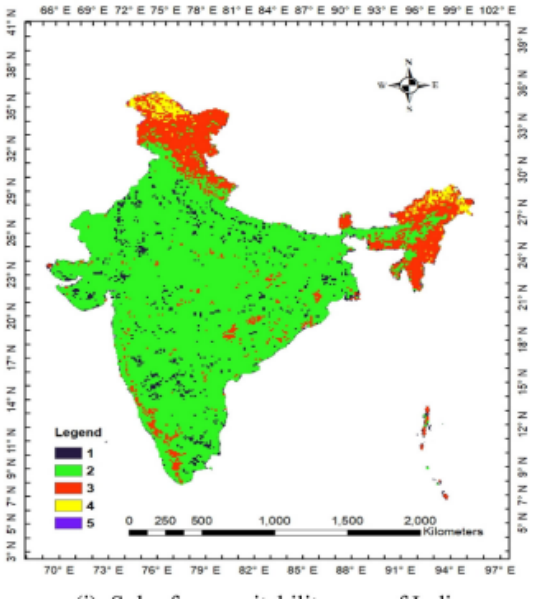


Figure 2.2: Optimal sites by level of importance by author [2]

2.2 Analytical Hierarchical Process

MCDMs have been extensively employed in literature for identifying optimal solar PV plant sites. Unlike machine learning techniques, which automatically learn biases without being explicitly programmed, MCDMs primarily focus on decision-making based on predefined criteria. These criteria are often ranked manually to guide decision-making processes.

AHP is a widely recognized MCDMs technique, as attested by numerous literary sources[1, 2, 8, 9]. AHP was pioneered by Prof. Thomas Saaty[13]. At its core, AHP emphasizes ranking criteria that influence the decision-making process.

The methodology of AHP can be distilled into three stages:

1. **Problem Definition and Hierarchy Creation:** Begin by clearly outlining the problem. For the context of this study, the objective is to assess the suitability of a location for a solar PV plant. A hierarchy is then defined

based on relevant criteria or factors, which in this instance might encompass aspects like elevation, slope, solar irradiance, land use, and land value.

2. **Sub-Criteria Classification:** The primary criteria can be further segmented into sub-criteria, enhancing the granularity of the hierarchical structure.

After establishing a hierarchy, it is essential to define the importance of criteria or factors relative to one another. This can be achieved using a technique known as pairwise comparison. This method involves comparing each factor with every other criterion, and the results of these comparisons can be stored in a matrix, termed the pairwise comparison matrix.

Once each factor's pairwise comparison with others is documented, the subsequent step is to determine the weights for each criterion. To do this, the matrix is first normalized. Subsequently, a weighted sum of the normalized criteria weights is computed to produce a score. From the normalized vector values, the Consistency Ratio (CR) is determined to validate the hierarchy's legitimacy.

The Consistency Ratio is a crucial metric that underpins the reliability of the decision-making process. A Consistency Ratio below 0.1 suggests that the weights generated can be deemed consistent and acceptable [13].

The Consistency Ratio (CR), Consistency Index (CI), and Random Consistency Index (RI) are interconnected. The formula to determine CR is:

$$CR = \frac{CI}{RI}$$

Here, RI serves as a reference value instrumental in gauging the consistency of pairwise comparisons. It offers a benchmark for verifying the attained consistency for the pre-defined hierarchy. Meanwhile, CI is derived from the eigenvalue of the

pairwise comparison matrix, and it is given by:

$$CI = \frac{\lambda_{\max} - n}{n - 1}$$

Where n represents the number of criteria under consideration, and $\lambda_{\max}$ is the largest eigenvalue.

2.3 Kohonen's Model

The Kohonen model, also known as the Kohonen neural network or Self-Organizing Map (SOM), is an unsupervised clustering algorithm. It was introduced by Kohonen et al. in 1982[14]. Typically, it is employed for clustering tasks. Notably, Chang et al. utilized it extensively to identify locations with high landslide susceptibility [10].

One of the primary objectives of the Kohonen model is dimensionality reduction. The aim is to generate a low-dimensional representation while retaining the inherent properties of the data. This is accomplished by associating each neuron with a weight vector that has the same dimension as the input data. These weights are iteratively aligned to match the distribution of the input data.

Initially, all the weight vectors of the neurons are initialized with random values drawn from a normal distribution. Iteratively, input vectors are selected from the training data. To compute the proximity of the input vector to the weighted vector, a distance or similarity metric, such as the Euclidean distance or cosine similarity, is employed. The Best Matching Unit (BMU) is the neuron whose weight vector most closely matches the input vector. Subsequently, the weights of the neurons are updated to align more closely with the selected input vector. This process continues iteratively until convergence.

Upon completion of the training phase, the Kohonen model produces a lower-

dimensional vector space representation of the input data. The number of neurons in the model signifies the number of classes or clusters intended for classification. It's worth noting that the Kohonen model can be supplied with vectorized GIS data spanning multiple criteria, as demonstrated by Chang et al. in their research on landslide susceptibility[10].

2.4 Autoencoder

The term Autoencoder has gained prominence in recent times due to advancements in deep learning and improvements in computational power of modern information systems. Nonetheless, it is important to acknowledge that the foundational principle of "learning representations" of data specific to a domain was established as early as 1980's. Thus the term might seem recent the foundation for the idea has existed for decades.

Rumelhart, Hinton, Williams, et al. introduced one of the earliest forms of networks in 1986, which can be viewed as the foundational milestone for autoencoders[15]. This study aim to learn representations vectors from the input vectors that represent important features of the task domain for the particular data. The network structure of the neuron-like functions with internal hidden units was proposed in the study for creating this representation.

2.4.1 Structure of the network

The authors proposed a network with that consisted of an input layer at the bottom and the output layer at the top with a intermediate laye. Conditions to this network stipulate that there can be no connections from the top of the network to the bottom. Although the connections can be skipped in the intermediate layer. The state at each of the neuron unit will be produced by two equations.

Where x_j would be the input to next unit j defined as a linear function that is the sum of all input y_i and their connection weights w_ij

$$x_j = \sum_i w_{ij} \times y_i$$

These units should be provided with biases unique to each intermidiate layer the the value 1. As the model is trying to solve non-linear problems. y_j to the next layer should be defined using a non-linear function.

$$y_j = \frac{1}{1 + e^{-x_j}}$$

2.4.2 Backpropagation Algorithm

The aim of the Backpropagation Algorithm is to find the weights w such that the input vector closely matches the output vector. For a finite number of cases, the total error can be computed by comparing the desired output to the actual output for each vector and then summing the errors of each individual case.

The total error E can be defined as the sum of the squared difference between the desired output d_j, c and the actual output y_j, c for each output neuron j and for each individual case c.

$$E = \frac{1}{2} \sum_c \sum_j (Y_{j,c} - d_{j,c})^2$$

To minimize the error (E) using gradient descent it is essential to calculate the partial derivatives of E with respect to each weight of the network using chain rule.

The backward pass starts with the output neuron where the actual output y_j is compared with the desired output d_j for each output neuron j.

$$\frac{\delta E}{\delta y_j} = y_j - d_j$$

Then the chain rule is applied to get the partial derivatives of previous layers neurons x_j.

$$\frac{\delta E}{\delta x_j} = \frac{\delta E}{\delta y_j} \cdot \frac{\delta y_j}{\delta x_j}$$

The partial derivatives calculated could be used to update the weights. The weights are updated in the amount proportional to the partial derivate using the learning rate parameter.

$$\Delta w = -\epsilon \frac{\delta \mathrm{E}}{\delta w}$$

In hindsight, the early neural networks, used for learning representations, laid the groundwork for the advent of deep learning in the 1980s. However, these networks presented several challenges that required solutions.

A primary concern centered around weight initialization. If all neurons in the network possessed identical weights, the system would fail to grasp the diverse functions crucial for accurate representation. To counter this, initializing the network with small random weights proved effective.

Additionally, the learning rate is an important hyperparameter. If set too high, the model might never converge, oscillating around the optimal solution. Conversely, a learning rate that's too low can lead to extremely slow convergence.

Even with an optimal learning rate, autoencoders aren't immune to challenges. A significant hurdle is the potential of getting trapped in a local minimum, mistakenly believing it has found the optimal solution when a better one might exist elsewhere.

With the surge in data volume, deep learning techniques and experiments

from the 2000s have made significant strides in addressing these issues.

2.5 Deep Autoencoder

While the early Autoencoder provided a foundation for neural networks that learn representations, the volume of information and complexity of data increased with the advent of internet causing the rise of deep learning to learn complex and intricate representations. A Deep Autoencoder is essentially based on similar principles discussed by Rumelhart et al. [15], but as the name suggest they are used to encode and decode using multiple hidden layers that gives them the ablility to learn hypotheses that can capture more complex patterns.

Hinton et al. proposed an deep Autoencoder structure in 2006 that can be used for dimensionality reduction as an alternative to Principle Component Analysis (PCA).

The autoencoder developed by Hinton el al was structurally segmented into three main components:

1. **Encoder:** It compresses the data into a lower-dimensional vector space. The encoder uses a neural network with transformational layers such as fully connected, convolutional, and dropout layers.

2. **Latent Space:** This represents the compressed version of the input vector produced by the encoder. Serves as a bottlenec which retains the core features of the data.

3. **Decoder:** This component of the neural network endeavors to reconstruct the original input vector. Its primary function lies in computing the reconstruction cost.

For the input to the encoder an ensemble of binary vectors such as images were modeled using Restricted Boltzmann machines (RBMs) to get a vector as a pretraining step. After pretraining the RBMs are unrolled and feed to the deep autoencoder for learning the representations in the latent space.

An inherent issues with Deep Autoencoders is the problem arising due to vanishing gradients during each iteration of backpropogation the network update the weights of the neuron units based on the calculated partial derivates as discussed in the previous section. The problem arises when the gradients become to small in deep networks causing the weights to not change halting the training process. Modern deep learning techniques like Batch Normalisation, Dropout and Residual blocks can helop solve the problem on vanishing gradients.

1. **Batch Normalisation:** Batch Normalisation is a method to solve vanishing gradient problem conjured by Ioffe et al.[16], Essentially it normalises the activations of each layer to have zero mean and variance which helps ensure that the activations don't have extremely high or low values . It helps to smoothen the error landscape, thus making for the gradients to converge.

2. **Dropout:** Dropout is a regularization technique where, random subset of neurons in layers are dropped during forward propogation to prevent overfitting. by randomly dropping neurons the network does't rely on a single path duing each training iteration, thus it also helps with the problem of vanishing gradients.

3. **Residual Blocks:** Using Batch Normalisation during the initial training iterations can cause gradient explosion problem. To solve this we can use residual blocks which skip certain connections creating a highway of activations skipping layers, allowing activations from a earlier part of the network

to be accessed by layers deeper in the network.

2.6 A previous study that utilizes Autoencoder for geospatial data

Ahmadlou et al. conducted comprehensive research on flood susceptibility in both Iran and India [3]. In their study, geospatial data, including layers such as slope, aspect, altitude, land use, and rainfall, served as determining criteria for the model [3].

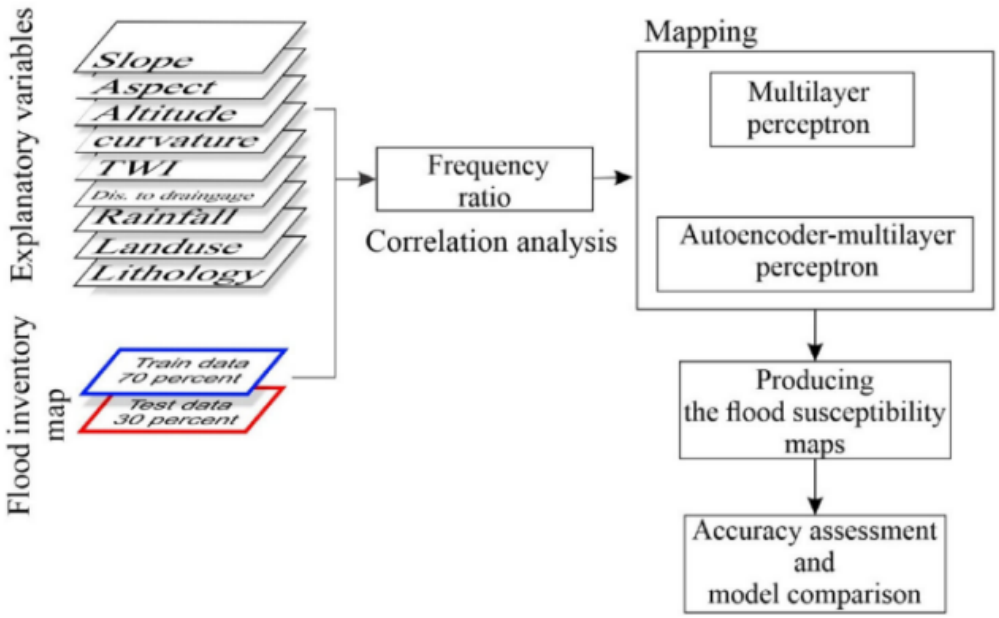


Figure 2.3: Flowchart of the model proposed by[3]

The study employed a hybrid model that combines a Multi-Layer Perceptron (MLP) and an Autoencoder. The data comprised geospatial features such as slope, curvature, aspect, altitude, rainfall, and land use. Hidden representations

of these features were derived using autoencoders.

The dataset was categorized into five classes representing flood risk levels: very low, low, moderate, high, and very high. Given the labeled nature of this dataset, they opted to use an MLP, which is a supervised learning method, to determine the model weights.

A MLP is a type of neural network that consists of one input layer which in the case study consist of the hidden representations learned from the autoencoder, one or multiple hidden layers connected to a output layer that predicts the class labels using the backpropogation algorithm along with a optimization algorithm like gradient descent to adjust model weight iteratively.

Chapter 3

Data

For the task at hand, a novel methodology is proposed in this project. The features considered for this task, can be subdivided into three classes:

1. **Terrain Information:** In this case, the elevation model, denoted by DEM, was readily available from the SRTM mission undertaken by USGS. This data, with a spatial resolution of 30m, is superior to the data used in previous studies by other authors. Additionally, the slope of the terrain is another crucial factor. The slope determines the amount of solar irradiance incident on the surface of the PV cells. Furthermore, a steep slope would not be suitable for a solar PV farm.

2. **Solar Irradiance and Atmospheric Conditions:** These were sourced from NREL. They play a pivotal role in identifying suitable locations. Specifically, a region with high solar irradiance, minimal overcast conditions, and high temperatures would be ideal.

3. **Other Features:** For this project, we also considered GIS layers from the OSM dataset. This includes features such as railways, roadways, land use, landmarks, and urban areas.

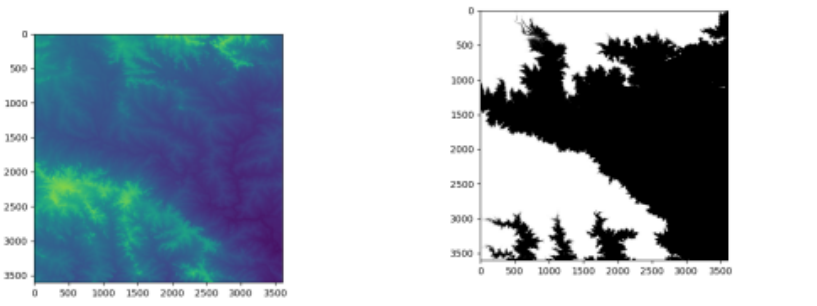
The primary objective here is to employ K-means clustering to create clusters and identify the most suitable ones based on locations that already contain a solar PV plant within the appropriate cluster. However, given the data's diversity and high dimensionality, it's essential to reduce its dimensionality and integrate our data from various sources before applying the clustering techniques.

3.1 Data Sources and Overview

For this project, we required multiple layers of GIS data that would serve as essential features for identifying suitable locations for solar farms. Terrain information is a pivotal feature for this study. To set up a solar farm, vast expanses of land with minimal elevation changes are necessary. Another critical attribute is solar irradiance, which is defined as the power potential generated from solar radiation incident on a specific location, measured in watts per square meter (W/m^2). Additional crucial factors to consider include land cost, population density, land use, and protected wildlife sanctuaries.

3.1.1 Terrain Data

The USGS) is an agency of the United States government that operates across disciplines such as geology, geography, and hydrology. The SRTM was undertaken in collaboration with National Aeronautical and Space Agency (NASA) to create DEM of the earth's surface. This effort produced two Digital Elevation Models available for research, with spatial resolutions of 1 arc-second (30 meters) and 3 arc-second (90 meters). For this study, we will be using the DEM model with a 1 arc-second spatial resolution[17].



(a) Elevation map

(b) Elevation values with a mask on the average

Figure 3.1: Elevation map for coordinates N 20' E 78

3.1.2 Solar Irradiance Data

The National Solar Irradiance Database (NSRDB) provides a comprehensive collection of solar irradiance data. This database, which is calculated on both hourly and half-hourly bases, is maintained by the NREL, the U.S. Department of Energy, and various other contributors[18].

Solar irradiance is characterized by three distinct measurements:

- **Global Horizontal Irradiance (GHI):** This refers to the total amount of solar radiation received per unit area on the Earth's surface. It is a cumulative measure that encompasses diffuse horizontal irradiance, ground-reflected radiation, and diffuse sky radiation.
- **Direct Normal Irradiance (DNI):** DNI indicates the amount of solar radiation received per unit area on a surface that is perpendicular to the sun rays incident on that surface.
- **Direct Horizontal Irradiance (DHI):** This measurement pertains to the solar radiation that is scattered from the sky, excluding the direct solar beam.

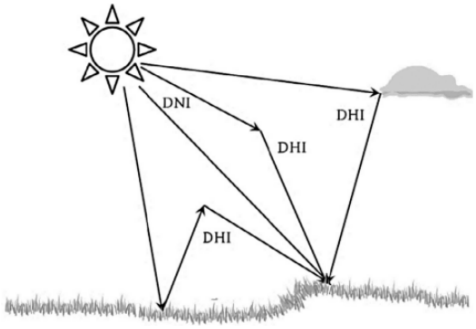
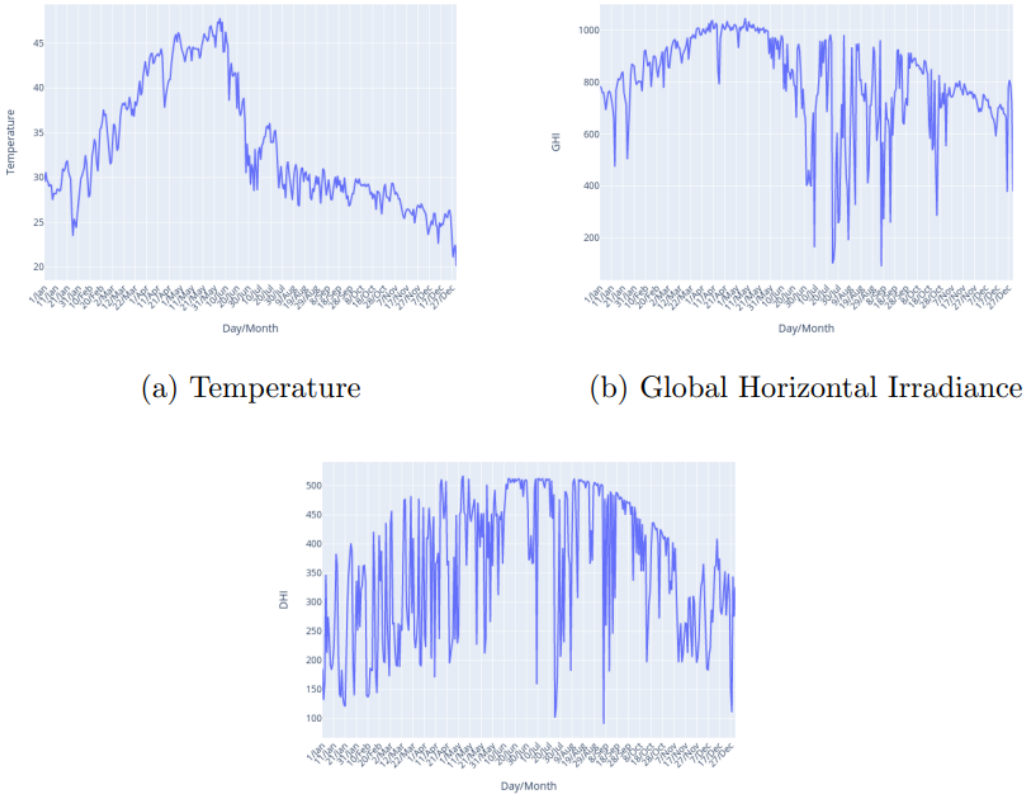


Figure 3.2: Solar Irradiance components[4]



(a) Temperature

(b) Global Horizontal Irradiance

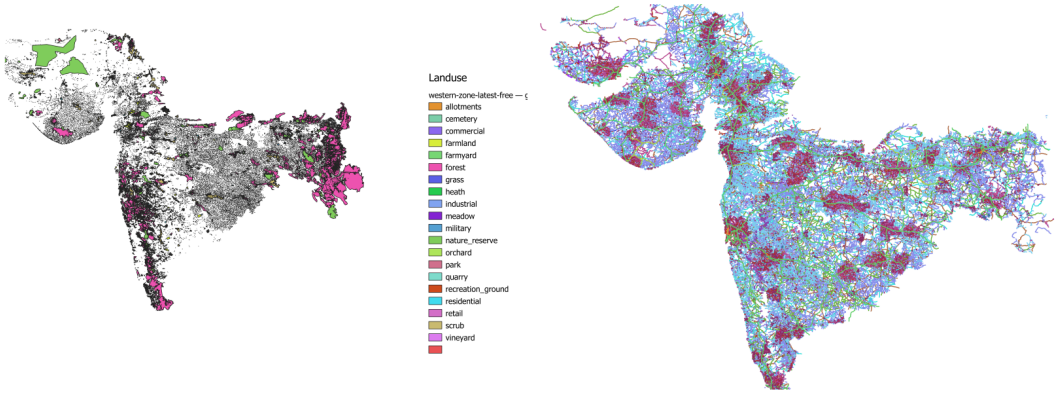
Figure 3.3: Solar irradiance data for co-ordinates for coordinates N 20' E 78'

3.1.3 Other Important Attributes

While solar irradiance and elevation are critical features to consider when planning the setup of a solar farm, there are numerous other factors that warrant attention. These include:

- **Financial Viability:** It's essential to assess whether there's sufficient demand for energy in the region to sustain a solar farm.
- **Environmental Impact:** Careful consideration must be given to the potential environmental repercussions of developing a solar plant, especially in ecologically sensitive regions.
- **Land-use Guidelines:** The designated or allowable uses of a land parcel can influence its suitability for solar farming.
- **Skilled Labor:** The availability of trained professionals and workers in the vicinity can have a bearing on the feasibility of the project.

OSM is an open-source collaborative project initiated in 2004, with the objective of creating free geographic data. Over the years, it has evolved into a global community-driven initiative. The maps and data generated by OSM collaborators can be leveraged for various attributes required for such projects [19].



(a) Land Use for Western India

(b) Road and Railways

Figure 3.4: Attributes for Western India

3.2 Data Modelling and Preprocessing

3.2.1 Terrain Data

The terrain data, available as a DEM model, boasts a spatial resolution of 30m and dimensions of 3601×3601. Each matrix value in this model represents the elevation. Consequently, each terrain data file encompasses an area of approximately 389,016 square kilometers. Although NREL does not provide direct slope data, it is possible to derive slope values from the elevation file. Mathematically, slope represents the rate of change of variable values. Given that we possess the adjacent elevation values and that there's a consistent 30-meter distance between these values, a function can be formulated to determine the slope values corresponding to the DEM data file.

Algorithm below showcases the calculation of slope matrix from the given elevation matrix.

Algorithm 1 Calculate Slope from Elevation matrix

Require: $matrix$, $spatial\_resolution$
Ensure: $slope\_matrix$
$rows, cols \leftarrow \text{shape}(matrix)$
$slope\_matrix \leftarrow$ zeros matrix of size $(rows, cols)$
for $i = 1$ to $rows - 2$ **do**
  for $j = 1$ to $cols - 2$ **do**
    $delta\_x \leftarrow matrix[i, j+1] - matrix[i, j-1]$
    $delta\_y \leftarrow matrix[i+1, j] - matrix[i-1, j]$
    $distance\_x \leftarrow 2 \times spatial\_resolution$
    $distance\_y \leftarrow 2 \times spatial\_resolution$
    $slope\_x \leftarrow \frac{delta\_x}{distance\_x}$
    $slope\_y \leftarrow \frac{delta\_y}{distance\_y}$
    $slope \leftarrow \sqrt{slope\_x^2 + slope\_y^2}$
    $slope\_matrix[i, j] \leftarrow slope$
  end for
end for
return $slope\_matrix$

By following the algorithm, one can obtain a matrix that provides an approximation of the slope at each cell in the DEM.

3.2.2 Solar Irradiance and Atmospheric Conditions Data

The atmospheric conditions data, fetched from NREL, is presented in the form of a time series for specific locations. For the model, we utilize the mean of the time series of various attributes (GHI, DHI, DNI, Clearsky DHI, Clearsky DNI, Clearsky GHI, Temperature, Relative Humidity). Essentially, we compute the mean of the time series for each location for which we have both elevation and slope data available.

3.2.3 Other GIS features

The GIS encompasses features such as railways, roadways, land use, landmarks, and urban areas. These features are represented as shapefiles, which comprise polygons mapped onto a coordinate system. These polygons facilitate the assignment of specific characteristics to designated areas. After utilizing the terrain data and atmospheric condition attributes for clustering, we will leverage the GIS features to develop a rule-based classifier. This classifier aims to subset land based on the polygon values from the GIS layers, as detailed in the associated table.

Table 3.1: Spatial Data Classification

Category	**Unsuitable area for the study**	**Suitable region for study**
Buildings	place_of_worship, parking, hotel, apartments, commercial, tower, train_station, hospital, house, university, dormitory, school, college, residential, industrial, bungalow, civic, attraction, transportation, storage_tank, cinema, stadium, service, depot, church, building, temple, fire_station, library, bank, public, retail, ruins, office, police, terrace, canteen, cafe, collapsed, mosque, restaurant, bus_station, memorial, construction, public_building, sports_centre, hostel, garage, supermarket, roof, toilets, hut, central_office, apartment, mall, wall, theatre, shed, detached, courthouse, clinic, workshop, fuel, cabin, post_office, marketplace, childcare, warehouse, community_centre, water_tower, fast_food, manufacture, blood_bank, events_venue, townhall, part, prep_school, guest_house, social_facility, gallery, arts_centre, kindergarten, yes,_hotel,_comm, hangar, yes, bar, ferry_terminal, greenhouse, fort, chapel, works, factory, shop, shelter, stage, city_gate, tomb, veterinary, lighthouse, car_wash, pumping_station, outdoor, motorcycle_parki, social_centre, clubhouse, motel, substation, food_court, stupa, government, pier, cowshed, aquarium, pharmacy	undefined, nan, vacant

Continued on next page

Table 3.1 – continued from previous page

Category	**Unsuitable area for the study**	**Suitable region for study**
landuse	commercial, retail, residential, recreation_groun, railway, salt_pond, grass, industrial, military, cemetery, plant_nursery, farmland, construction, basin, garages, quarry, reservoir, saltpond, brownfield, landfill, religious, allotments, yes, green, orchard, garden, village_green, education, farmyard, aquaculture, parking, used by school a	barren, meadow, vacant, greenfield
places	city, neighbourhood, locality, suburb, hamlet, village, town, island, yes, farm, plot, borough	None
railway	rail, platform, subway, miniature, monorail, construction, platform_edge, traverser, workshop, turntable, proposed	abandoned, disused
natural	water, forest, park	-
waterways	drain, river, stream, dam, weir, canal, dock	unclassified
road	residential, motorway, tertiary, service, trunk, path, secondary, primary, living_street, construction, footway, primary_link, track, motorway_link, trunk_link, steps, pedestrian, secondary_link, tertiary_link, bridleway, proposed, cycleway, elevator, bus_stop, services	-

Continued on next page

Table 3.1 – continued from previous page

Category	**Unsuitable area for the study**	**Suitable region for study**
points	switch, station, level_crossing, stop, traffic_signals, hotel, clinic, crossing, school, parking, mini_roundabout, bus_stop, hospital, restaurant, place_of_worship, motorway_junctio, attraction, ruins, studio, bank, theatre, fast_food, ice_cream, college, pharmacy, cafe, police, The Wall, pub, toilets, memorial, fuel, telephone, ferry_terminal, buffer_stop, water_tower, hostel, post_box, bus_station, lighthouse, atm, theme_park, fire_station, community_centre, post_office, fountain, picnic_site, works, cinema, marketplace, bar, monument, kindergarten, signal, railway_crossing, car_wash, doctors, townhall, viewpoint, clubhouse, guest_house, drinking_water, museum, battlefield, library, taxi, turning_circle, artwork, water_well, motel, fort, childcare, social_centre, tower, surveillance, flagpole, dentist, rest_area, vending_machine, public, passing_place, spa, parking_entrance, unclassified, bureau_de_change, platform, information, nightclub, chimney, bench, turning_loop, car_rental, internet_cafe, camp_site, events_venue, veterinary, gallery, crematorium, wayside_cross, waste_basket, stage, city_gate, garage, taxi_rank, arts_centre, hairdresser, border_control, swimming_pool, water_mill, milestone, well, fishing, bbq, leisure, lodge, boat_rental, water_tap, embassy, laundry, razor, bollard, power, toll_gate, street_lamp, photo_booth, pitch, water_point, barrier, prison, convenience, park, speed_camera, graveyard, bench	unclassified

In the table above, the 'Category' represents the layers of GIS, while the other columns contain the values associated with those layers. These values specify the type of location, represented as polygons, for the corresponding GIS layer.

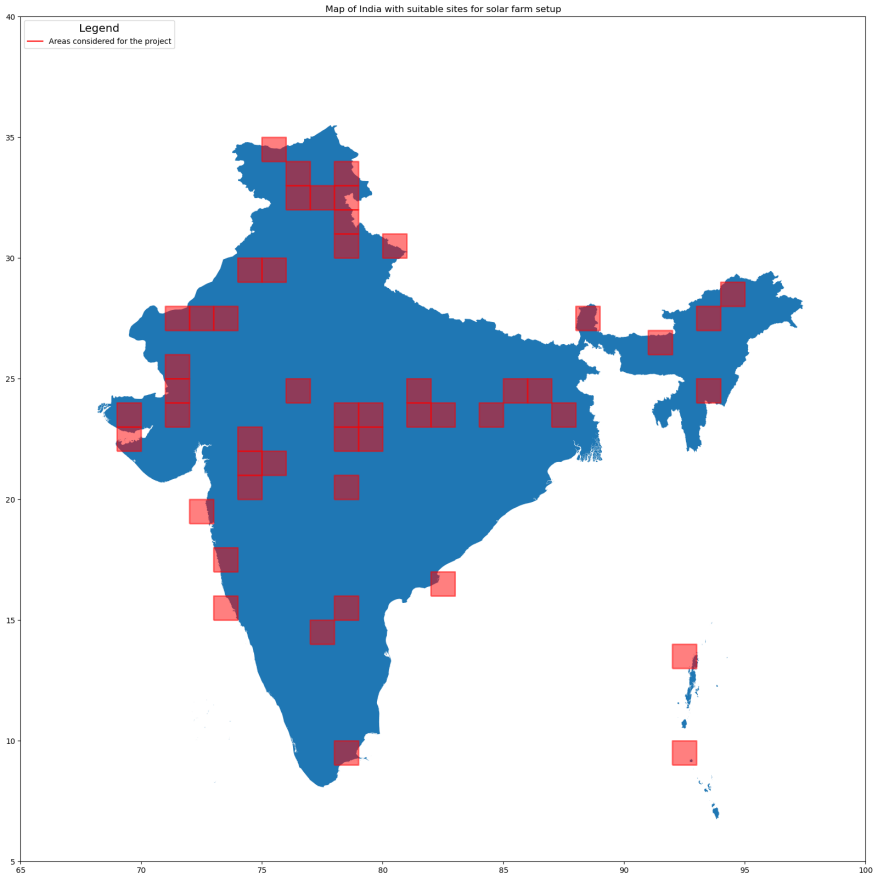


Figure 3.5: Map with areas highlighted that are considered for the study containing the Elevation and Slope data, solar irradiance and atmospheric conditions data and GIS information

Figure 3.5 highlights the geographic areas considered for this study. These specific regions were selected to capture a diverse range of locations for representation in the autoencoder. The chosen regions encompass a variety of geographical features, including tall mountains, desert landscapes, plains, plateaus, and islands.

3.2.4 Limitations of dataset

The data regarding slope and elevation features boasts impressive quality, with a spatial resolution of 30m. However, there's a gap in our matrix when it comes to atmospheric conditions. Instead of comprehensive coverage, this data is available solely as a time series for specific locations. Additionally, due to time and resource constraints, we only have 50 data samples to test the model. This limited dataset might present challenges in obtaining a holistic understanding, and as we move forward, it's imperative to optimally use this limited set for accurate model testing.

Chapter 4

Methodology

4.1 Overview

Autoencoders, as a neural network architecture, have been an essential subject of research in deep learning. A foundational understanding of autoencoders can be traced back to the works of Rumelhart et al. and Hinton et al [15] [20]. As presented in Chapter 2, these studies laid the groundwork for the current application and understanding of autoencoders. The Deep Autoencoder, emerging from these foundational studies, serves as the primary focus of this investigation.

4.2 AutoEncoder

4.2.1 Introduction

Data prepared from steps for Data Preprocessing and Modelling that were discussed in Chapter 3 would be used for developing the Autoencoder for this study. The Elevation data matrix, Slope data matric that was created using Algorithm 1, Atmospheric condition features include GHI, DHI, DNI, Clearsky_GHI, Clearsky_DHI, Clearsky_DNI, Temperature, Relative Humidity.

Initially, given that the data for Slope and Elevation are represented as matrices, they are to be processed using two deep autoencoders with similar architectures. This is done after normalizing these matrices. Once these intermediate representations or embeddings are generated, they will be concatenated and passed to another autoencoder. This process aims to yield the final embeddings, which will be utilized for clustering regions with similar geographical features. Since the data for solar irradiance and atmospheric conditions are singular values, they will be fed into the final autoencoder. This will be done after dimensionality reduction of the concatenated embeddings through a few initial layers.

4.2.2 Model Architecture

The data for dual deep autoencoders for the slope and elevation matrices are in from of matrixes with each value in matrix represention the elevation in meters and the slope gradient value at a spatial resolution of 30 meters.

Since the data is in the form of two-dimensional matrices, they can be fed through initial layers of convolutional layers for feature extraction.

Terminologies used in Convolutional Neural Networks (CNNs)

- **Convolution Operation:** At its core, convolution operations involve taking a filter of a size smaller than the input and slides it over the input to produce a feature maps.During the training phase of a Convolutional Neural Network (CNN), after the forward pass on the network the error is backpropogated and the filter/kernel values are adjusted based on partial derivatives of previous errors.

- **Kernel/Filters:** The kernels slide over the input matrix to produce the feature map. The number of kernels determine the amount of feature maps that can be produces. The dimensions of the kernels should match the

image dimension for it to convolve over the matrix.

- **Stride:** The value of stride determines the movement of filter over the input matrix. A stride value of 1 moves one pixel right at a time, for the value 2 it moves 2 pixels at the time. the value of stride determine the output matrix dimensions.

- **Padding:** Padding involves adding a vector border of zeroes around the input to control the spatial size of output matrix and to ensure that the border pixels of the input matrix are processed with equal weightage.

To gee the size of the output matrix after convolution operations we can you the below formula, where W_{in} the width of input matrix, F is the filter size and S is the stride value and W_{out} is the dimension of output matrix

$$W_{\text{out}} = \left\lfloor \frac{W_{\text{in}} - F + 2P}{S} + 1 \right\rfloor$$

In the autoencoder architecture, the encoder is responsible for generating embeddings through the application of convolution layers followed by dense layers. Consequently, the decoder seeks to upscale these embeddings. When considering the dense linear layers within the decoder, it is necessary to mirror the architecture of the dense layers present in the encoder. To recreate the spatial feature maps before they undergo the convolution process, the transposed convolution operation is used to learn the filters for deconvolutional layers in the decoder.

The transposed convolution formula to upscale the feature maps in decoder where W_{in} the width of input matrix, F is the filter size and S is the stride value and W_{out} is the dimension of output matrix

$$W_{\text{out}} = (W_{\text{in}} - 1) \times S - 2P + (F - 1) + 1$$

4.3 Clustering

With the embeddings created through thhe final autoencoder, K-means clustering an unsupervised clustering technique will be used to get cluster similar geospatial regions based on their euclidean distance and proximity to a centroid.

K-means clustering is a popular clustering technique that is widely recognised for its simiplicity, computational efficiency, interpretablity.

The K in K-means represents the number of clusters to be considered for the data in the latent space of the autoencoder. Initially, the centroids are randomly selected within this latent space. Consequently, each datapoint in the latent space is assigned to a cluster based on its proximity to the centroids. This proximity is measured using distance metrics such as Euclidean, Manhattan, and Minkowski distances. After assignment, the centroids are recalculated based on the data points assigned to them in the previous step. The process continues iteratively until convergence. Here, *"convergence"* typically signifies that the centroids remain relatively stable, with no significant shifts in their positions compared to earlier iterations.

The popularity of K-means doesn't overshadow its shortcomings. In unsupervised learning problems, the value of K is typically unknown. The Elbow curve method is one technique employed to determine the appropriate value for K. Iteratively, different values for K, usually ranging from 1 to 10, are used to generate multiple K-means models. For each model, the sum-of-squared distances from the points to their respective centroids are computed. These values are then plotted against the corresponding K values. The *Elbow* point on this curve is where the rate of decrease in the sum-of-squared distances sharply changes. Although this point is often considered a good estimate for K, it doesn't guarantee an optimal number of clusters. Another significant concern with K-means clustering is the algorithm's potential to get stuck in a local minimum.

4.4 Rule-Based Classifier

After clustering, we can analyze the results to identify suitable locations by looking for locations that have pre-existing solar PV plants, considering the geographic features of the datapoints in each cluster, etc.

After subsetting the data points based on the cluster, we can select the geospatial locations that correspond to the embeddings of those data points. These geospatial locations were chosen based on several factors including terrain features, slope, elevation, and atmospheric conditions such as solar irradiance, humidity, and temperature. However, even with these considerations, there is still a lack of clarity regarding which regions are suitable for solar PV plants. This is because we have not taken into account GIS layer information such as urban and rural centers, roadways, railways, and land use of the area. We will utilize the GIS information we have collected for these regions to further refine our selection.

The GIS data for the geospatial locations is available for 8 categories, for each category there are various attributes with their corresponding area polygons which can be considered suitable or unsuitable for this study.

Given the criteria for solar PV plant installation, it's essential to discriminate the suitability of geospatial locations based on their respective attributes. Using the landuse layer as an example, attributes like barren, vacant, and greenfield are suitable, while attributes such as city, neighbourhood, and suburb would be unsuitable for our specific objective. Based on above arguments we can make a rule based classifier that crops land that would be suitable for solar PV plant installation for the geospatial locations selected through the process of clustering.

- **Buildings:** This layer contains attributes associated with human development and therefore most of the attributes would contain area polygons that would be unsuitable with few expections like attributes vacant and

None in the dataset.

- **Landuse:** Based on the type of activity associated with the land, this layer offers valuable attributes such as 'barren', 'vacant', and 'greenfield' that align with the focus of our study.
- **Places:** Collection of human settlements like city, locality, industrial are contained in this layer.
- **Railways, Roads and waterways :** Transit systems are defined in this layer as area polygons around the roads and railways.
- **Natural:** Nature and wildlife reserves along with water bodies are highlighed in this layer.
- **Points:** Singular human-developed structures such as guest houses, water towers, and rest areas are presented in this layer.

A complete list detailing the suitability and unsuitability of these attributes is tabulated in Table 3.1.

Chapter 5

Experiments and results

This research explores the application of autoencoders and clustering to develop a novel approach for identifying prime locations suitable for solar infrastructure within India. Subsequent experiments apply the methodologies d iscussed to gather insights.

5.1 Experiments

5.1.1 Experiment Settings

The Experiments were run using three things in mind.

1. Optimization efforts were consistently aimed at maximizing the efficiency and accuracy of location predictions, all while taking into account the limited computational resources available

2. The visualisations should give a clear idea on suitable geolocations available for the regions under consideration.

3. The research questions raised in Chapter 1 are answered.

5.2 Results

Chapter 6

Conclusion

6.1 Limitations

6.2 Conclusion

Here you must zoom back out to evaluate the thesis. Mention limitations and weaknesses as well as contributions and possible future work.

References

[1] H. E. Colak, T. Memisoglu, and Y. Gercek, "Optimal site selection for solar photovoltaic (pv) power plants using gis and ahp: A case study of malatya province, turkey," *Renewable Energy*, vol. 149, pp. 565–576, Apr 2020. [Online]. Available: https://doi.org/10.1016/j.renene.2019.12.078 v, 2, 4, 6, 8

[2] S. Saraswat, A. K. Digalwar, S. Yadav, and G. Kumar, "Mcdm and gis based modelling technique for assessment of solar and wind farm locations in india," *Renewable Energy*, vol. 169, pp. 865–884, May 2021. [Online]. Available: https://doi.org/10.1016/j.renene.2021.01.056 v, 2, 7, 8

[3] M. Ahmadlou, A. Al-Fugara, A. R. Al-Shabeeb, A. Arora, R. Al-Adamat, Q. B. Pham, N. Al-Ansari, N. T. T. Linh, and H. Sajedi, "Flood susceptibility mapping and assessment using a novel deep learning model combining multilayer perceptron and autoencoder neural networks," *Journal of Flood Risk Management*, vol. 14, no. 1, Dec 2020. [Online]. Available: https://doi.org/10.1111/jfr3.12683 v, 16

[4] F. Vignola, J. Michalsky, and T. Stoffel, *Solar and Infrared Radiation Measurements*, 2nd ed. CRC Press, 2023. v, 21

[5] BP. (2021) Bp statistical review of world energy 2021. Available at : http://www.indiaenvironmentportal.org.in/files/file/bp1

[6] B. News. (2023) Cop26: India pm narendra modi pledges net zero by 2070. [Online]. Available: https://www.bbc.com/news/world-asia-india-59125143 1

[7] Reuters. (2022, Dec) India's solar boom reverses gas momentum, cements coal use: Maguire. [Online]. Available: https://www.reuters.com/world/india/indias-solar-boom-reverses-gas-momentum-cements-coal-use-maguire-2022-12-14/ 2

[8] H. Z. A. Garni and A. Awasthi, "Solar pv power plant site selection using a gis-ahp based approach with application in saudi arabia," *Applied Energy*, vol. 206, pp. 1225–1240, Nov 2017. [Online]. Available: https://doi.org/10.1016/j.apenergy.2017.10.024 2, 6, 8

[9] M. Zoghi, A. H. Ehsani, M. Sadat, M. j. Amiri, and S. Karimi, "Optimization solar site selection by fuzzy logic model and weighted linear combination method in arid and semi-arid region: A case study isfahan-iran," *Renewable and Sustainable Energy Reviews*, vol. 68, pp. 986–996, Feb 2017. [Online]. Available: https://doi.org/10.1016/j.rser.2015.07.014 2, 6, 8

[10] Z. Chang, Z. Du, F. Zhang, F. Huang, J. Chen, W. Li, and Z. Guo, "Landslide susceptibility prediction based on remote sensing images and gis: Comparisons of supervised and unsupervised machine learning models," *Remote Sensing*, vol. 12, no. 3, p. 502, Feb 2020. [Online]. Available: https://doi.org/10.3390/rs12030502 2, 10, 11

[11] A. Jain, R. Mehta, and S. K. Mittal, "Modeling impact of solar radiation on site selection for solar pv power plants in india," *International Journal*

of Green Energy, vol. 8, no. 4, pp. 486–498, May 2011. [Online]. Available: https://doi.org/10.1080/15435075.2011.576293 2

[12] S. Sindhu, V. Nehra, and S. Luthra, "Investigation of feasibility study of solar farms deployment using hybrid ahp-topsis analysis: Case study of india," *Renewable and Sustainable Energy Reviews*, vol. 73, pp. 496–511, Jun 2017. [Online]. Available: https://doi.org/10.1016/j.rser.2017.01.135 2

[13] T. L. Saaty, *What is the analytic hierarchy process?* Springer Berlin Heidelberg, 1988, pp. 109–121. [Online]. Available: https://doi.org/10.1007/978-3-642-83555-1_5 8, 9

[14] T. Kohonen, "Self-organized formation of topologically correct feature maps," *Biological Cybernetics*, vol. 43, no. 1, pp. 59–69, 1982. [Online]. Available: https://doi.org/10.1007/bf00337288 10

[15] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, Oct 1986. [Online]. Available: https://doi.org/10.1038/323533a0 11, 14, 30

[16] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015. [Online]. Available: https://arxiv.org/abs/1502.03167 15

[17] T. G. Farr and M. Kobrick, "Shuttle radar topography mission produces a wealth of data," *Eos Trans. AGU*, vol. 81, pp. 583–583, 2000. 19

[18] M. Sengupta, Y. Xie, A. Lopez, A. Habte, G. Maclaurin, and J. Shelby, "The national solar radiation data base (nsrdb)," *Renewable and Sustainable Energy Reviews*, vol. 89, pp. 51–60, 2018. 20

[19] O. contributors. (2017) Planet dump retrieved from https://planet.osm.org. [Online]. Available: https://www.openstreetmap.org 22

[20] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, Jul. 2006. [Online]. Available: https://doi.org/10.1126/science.1127647 30