Solar Potential Mapping

Ivașcu Andreea-Daria and Dima Alexandru

National University of Science and Technology POLITEHNICA Bucharest

1 The Need for Solar Farms

1.1 Global Demand

The transition to renewable energy sources has become a global priority due to the urgent need to address climate change, reduce greenhouse gas emissions and ensure long-term energy security. Among the various renewable options, solar energy stands out as one of the most abundant, clean and rapidly deployable resources. **Large-scale solar farms**, in particular, play **a crucial role in meeting rising energy demands** while promoting environmental sustainability.

1.2 The Importance of Solar Farms

The development of solar farms, however, is not simply a matter of placing photovoltaic panels on any available land. A **successful solar farm** requires **careful site selection** based on multiple factors, including solar radiation levels, land characteristics, accessibility and proximity to existing infrastructure such as transmission lines. In addition, ecological, agricultural and social considerations must be taken into account to minimize negative impacts and ensure sustainable integration into local communities.

1.3 Social and Economic Benefits

Solar farms contribute not only to environmental sustainability, but also to social and economic development. They create new job opportunities in construction, maintenance and technology, supporting local economies and communities. In the long term, solar energy can lower electricity costs, reduce dependence on imported fuels and increase energy security. Additionally, by promoting clean energy adoption, solar farms help build more resilient communities and encourage innovation in the green economy.

2 The Objectives of the Project

Through this project, satellite imagery from Sentinel-2 of a selected region over the span of 100 days will be analyzed in order to identify areas suitable for the development of solar farms. The analysis focuses on key environmental and infrastructural factors such as solar potential, land cover, terrain characteristics and proximity to essential networks (road access and electricity). In order to successfully predict areas suitable for solar farms, a ML model will be used that will classify and choose optimal locations, ensuring a scalable and data-drive approach to site selection.

3 Theoretical Background

Solar energy has emerged as one of the most promising renewable energy sources because it is clean, abundant and increasingly cost-effective. Although solar radiation can provide thousands of terawatts of technically accessible energy across the Earth's surface, photovoltaic power still accounts for a relatively small share of global electricity production. This highlights the importance of expanding solar infrastructure and carefully selecting suitable sites for development.

3.1 Placement of Solar Farms

In order to place solar farms as good as possible, technical, environmental, economic and social considerations need to be balanced. Suitable areas must have high solar potential, but other factors such as land cover, slope, elevation and accessibility to infrastructure also influence feasibility. At the same time, environmental and social acceptance need to be taken into consideration in order to avoid conflicts with natural habitats, water resources or cultural heritage.

3.2 How to Make Data Reliable?

By using satellite images, Geographic Information Systems (GIS) and remote sensing, the evaluation of these criteria is even more effective, as recent research shows. Through indices like NDVI, unsuitable areas can identified, covered by dense vegetation, therefore excluding major forest areas, but also, through other indices, water bodies or built-up areas can also be excluded. Multi-criteria decision making methods, such as the Analytic Hierarchy Process, also lead to creating more reliable data and are commonly applied to assign importance to evaluation factors and integrate them into a single suitability map. These methods, combined with spatial analysis, provide a systematic approach for identifying optimal areas for solar park development, reducing risks for investors and supporting sustainable energy planning, by making a more reliable and data-driven selection.

4 Methodology

Through satellite imagery, geographic analysis and machine learning techniques to identify suitable regions, this project offers a small scale data-driven approach to predict the best areas for solar farm development.

4.1 Data

The data used in this project was obtained from Sentinel-2 mission through the Copernicus Data Space Ecosystem (CDSE). Through a Python script, images corresponding to a specific area of interest (in this case, images of a plain area in Romania) were retrieved over the period of a year. To ensure sufficient coverage and variability, the first 100 products matching the query were downloaded for further analysis.

All the Sentinel-2 satellite data (L2A level) were downloaded in the .SAFE format, which contains spectral bands in .jp2 format. From these data, the **Red band (B04) – 665 nm** and **Near Infrared band (B08) – 842 nm** were extracted at **10 m resolution.** These bands were then used to compute vegetation indices and evaluate surface characteristics.

4.2 Data Pre-processing

The raw satellite images will be processed to extract relevant information for site suitability analysis. In this step, atmospheric correction, calculation of vegetation indices (NDVI) and derivation of additional spatial features such as slope, elevation and proximity to infrastructure will be included. The processed dataset will then be prepared as input features for the machine learning model.

NDVI Computation. The Normalized Difference Vegetation Index (NDVI) was computed using the standard formula: $\boldsymbol{NDVI} = \frac{(\boldsymbol{NIR}-\boldsymbol{RED})}{(\boldsymbol{NIR}+\boldsymbol{RED})}$, where NIR is the reflectance values in the Near Infrared band (B08), RED is the reflectance values in the Red band (B04). The median NDVI was calculated for each pixel to reduce seasonal effects, atmospheric variability and noise.

Cloud Fraction Computation. For each Sentinel-2 scene, the Scene Classification Layer (SCL) was used, which labels each pixel as cloud, shadow, water or land. From this, a cloud fraction map was derived, representing the proportion of observations affected by each pixel across the study period.

Topographic Data Integration. In addition to optical variables, topographic information from a Digital Elevation Model (DEM) was integrated. From the DEM, two variables were derived: elevation and slope, where slope is computed as the spatial derivative of elevation. These were used as additional predictors for the machine learning model.

4.3 Solar Score

An **initial solar potential score** was computed as a weighted combination of indicators (NDVI, cloud coverage, topography) for each pixel. Then, this score **served as the target variable during the training of the ML model**.

The weights of each indicator was calculated after the formulas from a known paper [1], therefore the NDVI indicator and the slope have a weight of 0.2 and the elevation of 0.1. Also, the quantity of solar radiation has a weight of 0.5, calculated using the NDVI and the cloud fraction.

4.4 Dataset Construction for the ML Model

To enable large-scale learning, the study area was divided into tiles of 1000 x 1000 pixels. From each tile, 5000 random pixels were sampled and for each pixel the following predictors were extracted: NDVI median, cloud fraction, slope and elevation. All these predictors formed the **feature matrix X**, while the solar score represented the **target vector y**. This ensured a balanced and representative training dataset.

4.5 Machine Learning Model

In order to predict the solar potential, a **Random Forest Regressor** was used, because it was very advantageous in modeling complex nonlinear relationships between variables, is robust to noise and missing values and also avoids overlearning through combining the results of multiple trees. Its main parameters were the *n_estimators = 100 (number of trees), max_depth = 15 (maximum tree depth)* and *n_jobs = -1 (parallel execution on all CPU cores)*.

The dataset was split into training (80%) and testing (20%) subsets. Model performance was then evaluated using the Mean Squared Error (MSE) metric.

4.6 Predictions

After training, the ML model was applied to the entire study area using the predictor values for each pixel. As a result, a solar potential map estimated via machine learning was obtained, saved in GeoTIFF format for integration into GIS.

4.7 Validation

The validation of the ML model was carried out by splitting the dataset in two: 80% for training the model and 20% for testing it. Therefore the Random Forest Regressor was trained on the training data and its performance was evaluated on the unseen test set. The Mean Squared Error (MSE) was used as the evaluation metric, providing a measure of how well the predicted solar scores matched the reference values.

5 Results

The machine learning and GIS-based workflow produced spatial maps indicating the potential suitability of different areas for solar energy development. The generated maps highlight both favorable areas for solar farms and unsuitable areas for this energetic development (areas with high slope, dense cloud cover or unsuitable land cover types). These maps show a spatially explicit visualization of solar energy potential, allowing quick identification of suitable areas for possible farm installation.

6 Possible Limitations

6.1 Data availability

Some datasets were not always available for the entire study area, which sometimes may reduce the accuracy of the results.

6.2 Resolution Constraints

The input data were resampled so that it could match the Sentinel-2 resolution (10m). However, the original DEM resolution may be lower, which can lead to uncertainty in slope and elevation estimates.

6.3 Temporal Coverage

NDVI and cloud frequency values are based on specific time periods and may not fully represent seasonal variations in vegetation or atmospheric conditions.

6.4 Simplified Assumptions

This approach of weighting the features using pairwise comparison and normalization, may not reflect necessarily real-world constraints such as land ownership or socio-economic factors.

References

1. Georgiou, A., & Skarlatos, D.: *Optimal site selection for sitting a solar park using multi-criteria decision analysis and geographical information systems*. Geoscientific Instrumentation, Methods and Data Systems, 5(2), 321-332(2016).