Small Project Submission Date

Study on Solar Energy Production

Alessandro Castelli 12246581

1 Introduction

In recent years, there has been a growing interest in renewable energies, at the expense of the use of fossil fuels. This shift in perspective reflects an increasing awareness of environmental impacts and the need for transitions towards more sustainable energy sources. Among the many options available, solar energy emerges as one of the most widely adopted solutions. Its popularity stems from its clean and sustainable nature, based on the abundance of solar radiation. This energy source plays a crucial role in reducing greenhouse gas emissions and contributing to the transition towards a greener and eco-friendly energy landscape.
The implementation of solar energy not only provides environmental benefits but also presents economic and technological opportunities. Solar technologies are constantly evolving, making the utilization of this resource more cost-effective and efficient. Investing in renewable energies, including solar energy, represents a crucial step towards a more sustainable and resilient energy future.
In this context, the main objective will be to develop predictive models capable of estimating energy generation levels based on various assumptions. The subsequent document will focus on this theme, proposing modifications, improvements, and further insights.

2 Outline

The project presented in the following pages focuses on the prediction of solar energy, utilizing datasets rich in diverse information. These data will be thoroughly explored in the subsequent paragraphs, providing a comprehensive overview of the context and variables involved in our analysis.

3 Two Main Datasets

In the following section, you will find the datasets that I have used and studied for this project. The phase of researching and analyzing the datasets was extensive, and below are the most interesting ones that I have identified.

3.1 Emhires Dataset

The EMHIRES dataset (European Meteorological derived High resolution RES generation time series), published by the `European Commission` [1], represents the first collec-

tion of European data on solar energy production made accessible to the wider public. This distinguished repository of information is derived from meteorological sources and provides various granularity options regarding geographical subdivisions.
Its availability at various levels of detail for different geographical areas allows users to analyze and comprehend solar energy production in Europe in a thorough manner. This approach provides scholars, professionals, and enthusiasts in the field with a broad spectrum of data, facilitating the conduct of detailed analyses and enabling a more comprehensive understanding of dynamics related to solar energy in the European context.

3.1.1 Territories and Time Periods Covered by the EMHIRES Dataset

EMHIRES provides time series on renewable energy generation, focusing on solar energy, for the European Union and neighboring countries. The time series are detailed on an hourly basis and offer various levels of aggregation, such as by country, bidding zone in the electricity market, and NUTS 1 and NUTS 2 levels according to EUROSTAT classification.
The main goal of EMHIRES is to enable users to assess the impact of meteorological and climatic variations on solar energy production in Europe. The time series cover the period 1986-2015, without considering changes in installed solar capacity.

3.1.2 Countries Included in the Dataset

The dataset includes data from the following countries:

- AT = AUSTRIA
- BE = BELGIUM
- BG = BULGARIA
- CH = SWITZERLAND
- CY = CYPRUS
- CZ = CZECH REPUBLIC
- DE = GERMANY
- DK = DENMARK
- EE = ESTONIA
- ES = SPAIN
- FI = FINLAND
- FR = FRANCE
- EL = GREECE
- HR = CROATIA
- HU = HUNGARY
- IE = IRELAND
- IT = ITALY
- LT = LITHUANIA
- LU = LUXEMBOURG
- LV = LATVIA
- NL = NETHERLANDS
- NO = NORWAY
- PL = POLAND
- PT = PORTUGAL

- RO = ROMANIA
- SE = SWEDEN
- SI = SLOVENIA
- SK = SLOVAKIA
- UK = UNITED KINGDOM

3.1.3 Details on How EMHIRES Data Was Collected

The general approach to convert solar resources into energy generation involves the conversion of satellite-based radiation data using the PVIGS model [2]. Initially, meteorological data is collected, and subsequently, these data are converted into theoretical potential, i.e., the solar electricity generation in each area expressed in kW generated per kW peak of a typical Photovoltaic (PV) system. This means that the solar energy value in the dataset is expressed in terms of the ratio between the energy actually produced and the energy that could be produced with a nominal power of 1 kW. This ratio is called the capacity factor or performance ratio and depends on various factors such as solar radiation, temperature, orientation and tilt of panels, losses, and system efficiency. Finally, to obtain power generation, the installed capacity of each region is calculated. The time series are then corrected with the actual TSO (Transmission System Operator) generation and statistically validated for power system analysis, assessing power peaks and variations, duration curves, and capacity factors.
This process implies that in each cell of the dataset, there will be a value between 0 and 1 representing the ratio between the energy actually produced and the energy that could be produced.
Furthermore, the PVGIS model uses solar radiation data from sources such as the CM SAF SARAH (Solar surface Radiation Heliosat) solar radiation product[1], with a spatial resolution of 3 arc-minutes, and considers various factors such as low-angle reflectivity, temperature, and cooling of photovoltaic modules to calculate electricity generation. This approach provides a homogeneous methodology for simulating solar energy production at the national and regional levels in Europe, avoiding the use of artificial correction factors and reducing uncertainty in the results.

3.1.4 How Data is Structured in EMHIRES

The dataset is structured by rows, totaling 262,968 observations in total. Each column in the dataset represents a country, while each row corresponds to a specific hour, starting from January 1, 1986, until 2015, providing a total of 30 years of data.
The dataset structure can be visualized in Figure 1. Each data point in the dataset is identified by a combination of country and hour (the hour being implicit information).

[1]The CM SAF SARAH (Solar surface Radiation Heliosat) is a climate data product belonging to the Climate Monitoring Satellite Application Facility (CM SAF), a part of the European Organization for the Exploitation of Meteorological Satellites (EUMETSAT). The CM SAF SARAH dataset focuses on surface solar radiation and includes information on direct (horizontal and normal direct) and diffuse radiation under clear sky conditions.

	AT	BE	BG	CH	CY	CZ	DE	DK	EE	ES	...	LV	NL	NO	\
0	0.0	0.0	0.0	0.0	0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	
1	0.0	0.0	0.0	0.0	0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	
2	0.0	0.0	0.0	0.0	0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	
3	0.0	0.0	0.0	0.0	0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	
4	0.0	0.0	0.0	0.0	0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	
...	...	...	...	...	..	...	...	...	...	...	...	...	...	...	
262963	0.0	0.0	0.0	0.0	0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	
262964	0.0	0.0	0.0	0.0	0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	
262965	0.0	0.0	0.0	0.0	0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	
262966	0.0	0.0	0.0	0.0	0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	
262967	0.0	0.0	0.0	0.0	0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	

	PL	PT	RO	SI	SK	SE	UK
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0
...	...	...	...	...	...	...	...
262963	0.0	0.0	0.0	0.0	0.0	0.0	0.0
262964	0.0	0.0	0.0	0.0	0.0	0.0	0.0
262965	0.0	0.0	0.0	0.0	0.0	0.0	0.0
262966	0.0	0.0	0.0	0.0	0.0	0.0	0.0
262967	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Figure 1: EMHIRES dataset example

3.2 Monthly Electricity Statistics

The data contained in this dataset has been collected by the International Energy Agency (IEA) [3]. These data include information on energy in various countries from 2010 to 2022. Energy production is measured in gigawatt-hours (GWh) and covers various energy products, including solar energy. You can get an idea of the structure of this dataset from Figure 2.

3.2.1 Dataset Features

The dataset contains various features, which are as follows:

COUNTRY: Name of the country

CODE_TIME: A code representing the month and year (e.g., JAN2010 for January 2010)

TIME: The month and year in a more human-readable format (e.g., January 2010)

	COUNTRY	CODE_TIME	TIME	YEAR	MONTH	MONTH_NAME
0	Australia	JAN2010	January 2010	2010	1	January
1	Australia	JAN2010	January 2010	2010	1	January
2	Australia	JAN2010	January 2010	2010	1	January
3	Australia	JAN2010	January 2010	2010	1	January
4	Australia	JAN2010	January 2010	2010	1	January
...	...	...	...	...	...	...
181910	United States	DEC2022	December 2022	2022	12	December
181911	United States	DEC2022	December 2022	2022	12	December
181912	United States	DEC2022	December 2022	2022	12	December
181913	United States	DEC2022	December 2022	2022	12	December
181914	United States	DEC2022	December 2022	2022	12	December

	PRODUCT	VALUE	DISPLAY_ORDER
0	Hydro	990.728000	1
1	Wind	489.469000	2
2	Solar	49.216000	3
3	Geothermal	0.083000	4
4	Total combustible fuels	19289.730000	7
...	...	...	...
181910	Non-renewables	292417.548132	23
181911	Others	8017.840957	24
181912	Other renewables aggregated	6133.265943	25
181913	Low carbon	146425.474534	26
181914	Fossil fuels	223357.219650	27

	yearToDate	previousYearToDate	share
0	1.647189e+04	NaN	0.047771
1	4.940909e+03	NaN	0.019744
2	9.082380e+02	NaN	0.002373
3	9.960000e-01	NaN	0.000004
4	2.143030e+05	NaN	0.930108
...	...	...	...
181910	3.355042e+06	3.320634e+06	0.791164
181911	5.393606e+04	4.899452e+04	0.021693
181912	7.100997e+04	7.242158e+04	0.016594
181913	1.749805e+06	1.670531e+06	0.396168
181914	2.583925e+06	2.542138e+06	0.604315

Figure 2: IEA dataset example

YEAR: The year of the data point

MONTH: The month of the data point as a number (1-12)

MONTH_NAME: The month of the data point as a string (e.g., January)

PRODUCT: The type of energy product (e.g., hydropower, wind, solar)

VALUE: The amount of electricity generated in gigawatt-hours (GWh)

DISPLAY_ORDER: The order in which products should be displayed

yearToDate: The amount of electricity generated for the current year up to the current month in GWh

previousYearToDate: The amount of electricity generated for the previous year up to the current month in GWh

share: The share of the product in the total electricity production for the country in decimal format

3.2.2 Dataset Statistics

IEA has compiled accessible statistics on the webpage https://www.iea.org/data-and-statistics/data-tools/monthly-electricity-statistics. These statistics provide a comprehensive overview of how energy production has changed over the years and offer detailed information on the most utilized renewable energy sources.
It is evident that, in this historical period, solar energy is one of the prominent renewable energies, as can be observed by analyzing the data presented in the statistics. I do not include the images available on that website as the rights related to those graphics are reserved.

4 Other Datasets

The datasets previously presented are considered more "complete" and reliable as they have been published by European authorities or reputable companies. However, to conduct my study, I have also employed other datasets that may have less detailed or less reliable documentation. Despite this, I believe they are still correct since, upon analyzing the data, it is evident that they exhibit a certain degree of coherence. I will explain in more detail the reason for this choice later.

4.1 World Weather Repository

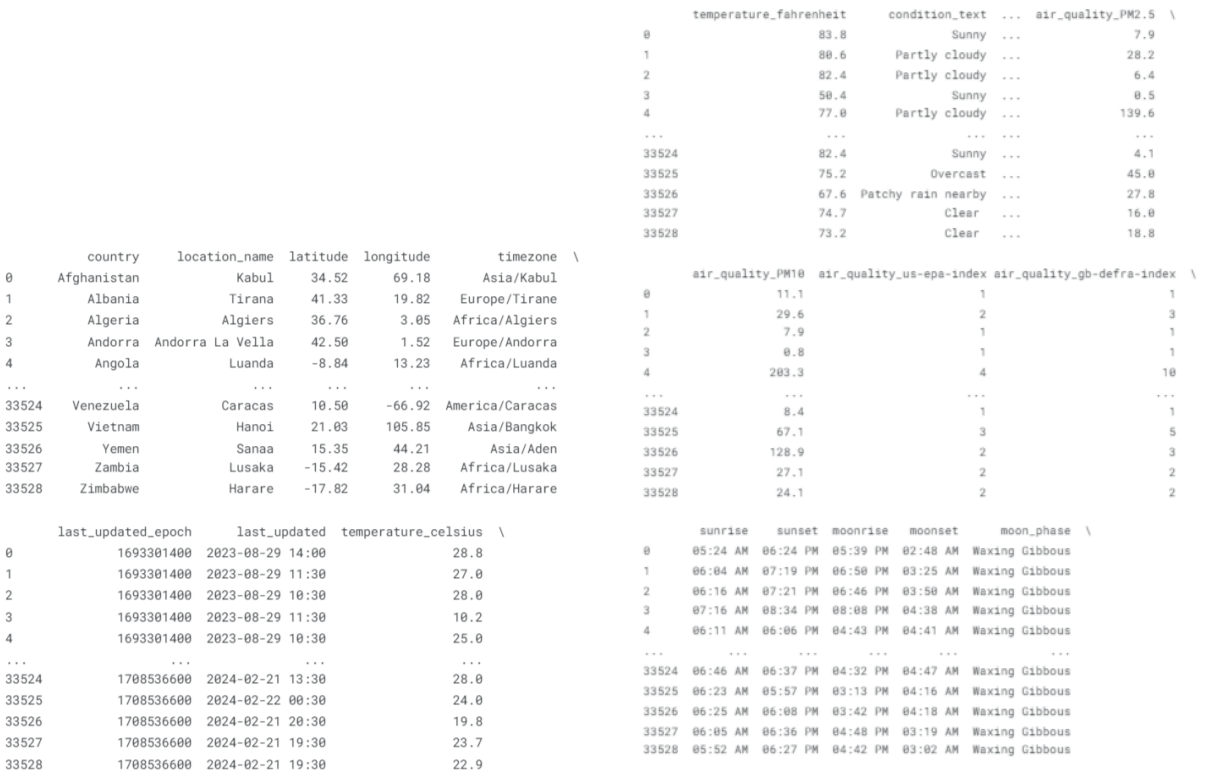
This dataset [5] provides daily meteorological information for capitals worldwide. Unlike forecast data, this dataset offers a complete set of features reflecting current weather conditions globally. It started collecting data from August 29, 2023.
The dataset includes over 40 features, including temperature, wind, pressure, precipitation, humidity, visibility, air quality measurements, and more. This extensive set of data is valuable for analyzing global meteorological patterns, exploring climate trends, and understanding relationships between different weather parameters.
The dataset features are as follows:

- **country:** Country of meteorological data
- **location_name:** Name of the location (city)
- **latitude:** Latitude coordinate of the location
- **longitude:** Longitude coordinate of the location
- **timezone:** Timezone of the location
- **last_updated_epoch:** Unix timestamp of the last data update
- **last_updated:** Local time of the last data update
- **temperature_celsius:** Temperature in degrees Celsius
- **temperature_fahrenheit:** Temperature in degrees Fahrenheit
- **condition_text:** Description of

weather conditions

- **wind_mph:** Wind speed in miles per hour
- **wind_kph:** Wind speed in kilometers per hour
- **wind_degree:** Wind direction in degrees
- **wind_direction:** Wind direction as a 16-point compass
- **pressure_mb:** Pressure in millibars
- **pressure_in:** Pressure in inches
- **precip_mm:** Amount of precipitation in millimeters
- **precip_in:** Amount of precipitation in inches
- **humidity:** Humidity in percentage
- **cloud:** Cloud cover in percentage
- **feels_like_celsius:** Feels-like temperature in degrees Celsius
- **feels_like_fahrenheit:** Feels-like temperature in degrees Fahrenheit
- **visibility_km:** Visibility in kilometers
- **visibility_miles:** Visibility in miles
- **uv_index:** UV index
- **gust_mph:** Wind gust in miles per hour
- **gust_kph:** Wind gust in kilometers per hour
- **air_quality_Carbon_Monoxide:** Air quality measurement: Carbon Monoxide
- **air_quality_Ozone:** Air quality measurement: Ozone
- **air_quality_Nitrogen_dioxide:** Air quality measurement: Nitrogen Dioxide
- **air_quality_Sulphur_dioxide:** Air quality measurement: Sulphur Dioxide
- **air_quality_PM2.5:** Air quality measurement: PM2.5
- **air_quality_PM10:** Air quality measurement: PM10
- **air_quality_us-epa-index:** Air quality measurement: US EPA Index
- **air_quality_gb-defra-index:** Air quality measurement: GB DEFRA Index
- **sunrise:** Local time of sunrise
- **sunset:** Local time of sunset
- **moonrise:** Local time of moonrise
- **moonset:** Local time of moonset
- **moon_phase:** Current moon phase
- **moon_illumination:** Percentage of moon illumination

You can see some rows of the dataset in Figure 3



	country	location_name	latitude	longitude	timezone
0	Afghanistan	Kabul	34.52	69.18	Asia/Kabul
1	Albania	Tirana	41.33	19.82	Europe/Tirane
2	Algeria	Algiers	36.76	3.05	Africa/Algiers
3	Andorra	Andorra La Vella	42.50	1.52	Europe/Andorra
4	Angola	Luanda	-8.84	13.23	Africa/Luanda
...	...	...	...	...	...
33524	Venezuela	Caracas	10.50	-66.92	America/Caracas
33525	Vietnam	Hanoi	21.03	105.85	Asia/Bangkok
33526	Yemen	Sanaa	15.35	44.21	Asia/Aden
33527	Zambia	Lusaka	-15.42	28.28	Africa/Lusaka
33528	Zimbabwe	Harare	-17.82	31.04	Africa/Harare

	last_updated_epoch	last_updated	temperature_celsius
0	1693301400	2023-08-29 14:00	28.8
1	1693301400	2023-08-29 11:30	27.0
2	1693301400	2023-08-29 10:30	28.0
3	1693301400	2023-08-29 11:30	10.2
4	1693301400	2023-08-29 10:30	25.0
...	...	...	...
33524	1708536600	2024-02-21 13:30	28.0
33525	1708536600	2024-02-22 00:30	24.0
33526	1708536600	2024-02-21 20:30	19.8
33527	1708536600	2024-02-21 19:30	23.7
33528	1708536600	2024-02-21 19:30	22.9

	temperature_fahrenheit	condition_text	...	air_quality_PM2.5
0	83.8	Sunny	...	7.9
1	80.6	Partly cloudy	...	28.2
2	82.4	Partly cloudy	...	6.4
3	50.4	Sunny	...	0.5
4	77.0	Partly cloudy	...	139.6
...	...	...	...	...
33524	82.4	Sunny	...	4.1
33525	75.2	Overcast	...	45.0
33526	67.6	Patchy rain nearby	...	27.8
33527	74.7	Clear	...	16.0
33528	73.2	Clear	...	18.8

	air_quality_PM10	air_quality_us-epa-index	air_quality_gb-defra-index
0	11.1	1	1
1	29.6	2	3
2	7.9	1	1
3	0.8	1	1
4	203.3	4	10
...	...	...	...
33524	8.4	1	1
33525	67.1	3	5
33526	128.9	2	3
33527	27.1	2	2
33528	24.1	2	2

	sunrise	sunset	moonrise	moonset	moon_phase
0	05:24 AM	06:24 PM	05:39 PM	02:48 AM	Waxing Gibbous
1	06:04 AM	07:19 PM	06:50 PM	03:25 AM	Waxing Gibbous
2	06:16 AM	07:21 PM	06:46 PM	03:50 AM	Waxing Gibbous
3	07:16 AM	08:34 PM	08:08 PM	04:38 AM	Waxing Gibbous
4	06:11 AM	06:06 PM	04:43 PM	04:41 AM	Waxing Gibbous
...	...	...	...	...	...
33524	06:46 AM	06:37 PM	04:32 PM	04:47 AM	Waxing Gibbous
33525	06:23 AM	05:57 PM	03:13 PM	04:16 AM	Waxing Gibbous
33526	06:25 AM	06:08 PM	03:42 PM	04:18 AM	Waxing Gibbous
33527	06:05 AM	06:36 PM	04:48 PM	03:19 AM	Waxing Gibbous
33528	05:52 AM	06:27 PM	04:42 PM	03:02 AM	Waxing Gibbous

Figure 3: World Weather Repository dataset example

4.2 Norway meteorological data

The dataset [4] represents a collection of precompiled and pre-processed meteorological data obtained from the meteorologisk institutt archive in Oslo, Norway, through its Frost API (https://frost.met.no/index.html). Data is collected from 55 meteorological stations distributed throughout Norway from January 1, 2010, to December 31, 2021. Each record in the dataset includes the station's identifier and geographical coordinates, date breakdown, and various observations such as maximum and average air temperature for the specific day, maximum and average wind speed, maximum relative humidity, and the sum of precipitation for the specific day. It's important to note that some values may be missing (`NaN`) as some stations only record some of the mentioned metrics.
To ensure consistency and ease of use of the dataset, you can proceed with the removal of unnecessary or redundant information. Additionally, you can further expand the dataset by incorporating additional variables or metrics of interest for a specific study.
The features present in this dataset are:

- Unnamed: 0
- sourceId

- latitude
- longitude
- max(air_temperature P1D)
- max(relative_humidity P1D)
- max(wind_speed P1D)
- mean(air_temperature P1D)
- mean(relative_humidity P1D)
- mean(wind_speed P1D)
- sum(precipitation_amount P1D)
- day
- month
- year

To see an example of how this dataset is structured, you can refer to Figure 4

5 Cleaning and Analysis

After a lengthy phase of identifying datasets, during which many were discarded from an initial group, I initiated the subsequent phase of cleaning and in-depth analysis (you can refer to the code here).
During this process, I carefully examined the datasets, identified the most relevant features, and addressed the handling of missing values, replacing them with appropriate values. Additionally, I observed the overall behavior of the datasets over time.
A crucial aspect of my study was exploring the existence of any trends, particularly checking if there was a consistent increase in energy production during summer months compared to winter months. For this analysis, I created various visualizations. A result contradicting this expectation could indicate a potential issue with data quality.
During this phase, I paid particular attention to ensuring the integrity and reliability of the data, aiming to obtain meaningful results for a correct interpretation of patterns and relationships in the datasets.
Figures 5 and 6 clearly show how the graphs exhibit the expected behavior.

	Unnamed: 0	sourceId	latitude	longtitude
0	0	SN100	61.134900	12.503900
1	1	SN1135	58.990000	11.540800
2	2	SN1151800	50.100278	14.255556
3	3	SN15262	61.674000	8.368500
4	4	SN1531000	45.473056	28.032222
...	...	...	...	...
237624	237624	SN50110	60.337200	6.217500
237625	237625	SN98978	70.335700	28.964800
237626	237626	SN299300	60.301111	19.131389
237627	237627	SN50110	60.337200	6.217500
237628	237628	SN98978	70.335700	28.964800

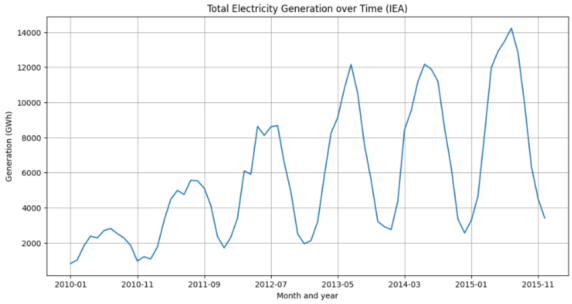
	max(air_temperature P1D)	max(relative_humidity P1D)
0	NaN	NaN
1	NaN	NaN
2	NaN	NaN
3	NaN	NaN
4	NaN	NaN
...	...	...
237624	NaN	NaN
237625	-11.25	89.0
237626	NaN	NaN
237627	NaN	NaN
237628	-7.00	94.0

	max(wind_speed P1D)	mean(air_temperature P1D)
0	NaN	NaN
1	NaN	NaN
2	NaN	NaN
3	NaN	NaN
4	NaN	NaN
...	...	...
237624	NaN	NaN
237625	2.6	-12.45
237626	NaN	NaN
237627	NaN	NaN
237628	NaN	-10.90

	mean(relative_humidity P1D)	mean(wind_speed P1D)
0	NaN	NaN
1	NaN	NaN
2	NaN	NaN
3	NaN	NaN
4	NaN	NaN
...	...	...
237624	NaN	NaN
237625	89.0	1.8
237626	NaN	NaN
237627	NaN	NaN
237628	91.0	NaN

	sum(precipitation_amount P1D)	day	month	year
0	0.4	1	1	2010
1	NaN	1	1	2010
2	NaN	1	1	2010
3	NaN	1	1	2010
4	NaN	1	1	2010
...	...	...	...	...
237624	NaN	29	12	2021
237625	NaN	29	12	2021
237626	NaN	30	12	2021
237627	NaN	30	12	2021
237628	NaN	30	12	2021

Figure 4: Norway-Meteorological-Dataset dataset example



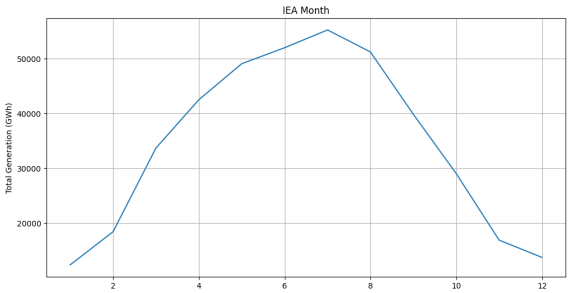


Figure 5: IEA plot

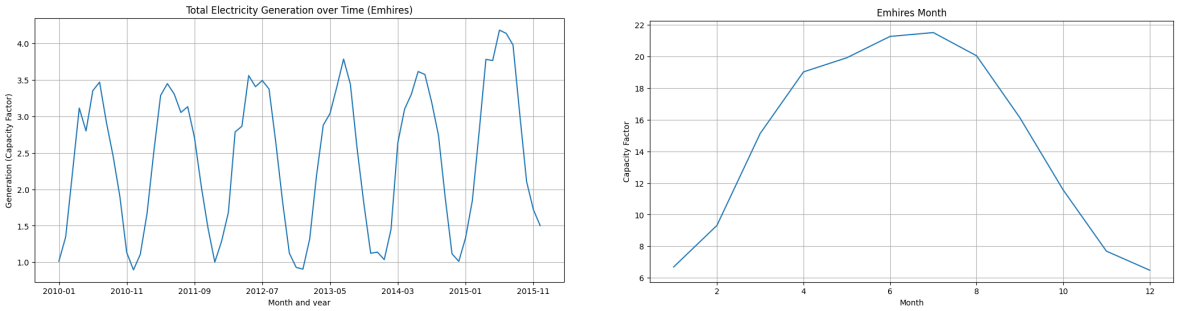


Figure 6: EMHIRES plot

6 Models

6.1 EMHIRES Model

After completing the phase of studying and cleaning the datasets, I proceeded with the development of regression models. I started using the EMHIRES dataset, and the code is available EMHIRES Model.
After an initial data pre-processing phase, I created several regression models, starting with simpler approaches and gradually advancing to more complex models, such as Recurrent Neural Networks (RNN). The EMHIRES dataset is of considerable size, allowing for good performance even with a limited number of epochs.
After training, I managed to achieve the following results:

- `Training Loss: 0.03785869851708412`
- `Validation Loss: 0.03948716109068664`
- `Test Loss: 0.03929314762353897`

These values are quite good and indicate that the model has adapted satisfactorily to the data.

6.2 IEA Monthly Electricity Statistics

For the IEA dataset, I followed the same process as described earlier (IEA Code). Initially, I performed a preprocessing phase on the dataset, including normalization, removal of unnecessary features, and one-hot encoding. Subsequently, I created a regression model using an RNN. Since the dataset is significantly smaller than the previous one, I used a higher number of epochs during training.

After the training phase, I obtained the following results for the loss metrics:

- `Training Loss: 0.012684706598520279`
- `Validation Loss: 0.14131624661014952`
- `Test Loss: 0.08876131474971771`

These values indicate that the model has adapted satisfactorily to the data.
I also calculated some model evaluation metrics:

- **R-squared (r2):** Represents the proportion of variance explained by the model, obtaining a value of `0.9294676281171568`, indicating good predictive ability.
- **Mean Squared Error (MSE):** Measures the average of the squared errors and is `0.07650821274623916`.

- **Mean Absolute Error (MAE):** Represents the average of absolute errors and has a value of `0.20185678746688712`.

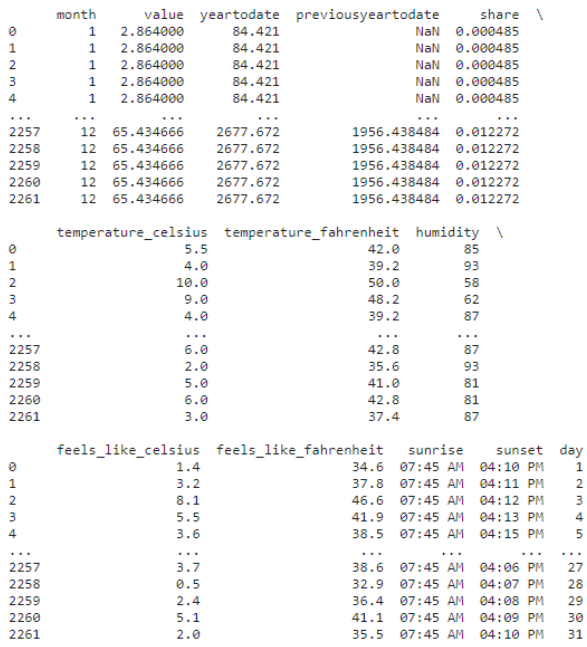
These metrics reflect good accuracy of the model in predicting IEA dataset's data. The high R-squared value suggests that a significant portion of the variance is explained by the model, and the low MSE and MAE values indicate good adaptability to the data.

6.3 Mixed Model

After training the main datasets, it was decided to also consider day and night hours to check if adding information about sunrise and sunset times could lead to more accurate performance. To test this hypothesis, I combined the IEA Dataset with the World Weather Repository (which contains sunrise and sunset times for some countries, and you can find the code here).
I focused only on the country "Austria" since the World Weather Repository dataset does not contain all the countries present in the IEA dataset. However, it is reasonable to assume that if better values are obtained for one country, the same reasoning can be extended to all the other available countries.
From Figure 7, you can see how the dataset looks after merging and eliminating non-significant columns.
On this dataset, I performed one-hot encoding for the dates and then proceeded to train the model.
The obtained metrics are as follows (with the same model parameters):

- `Training Loss: 0.0004668822803068906 (Figure 8a)`
- `Validation Loss: 0.021887180635916294 (Figure 8b)`
- `Test Loss: 0.08139250427484512`
- `R-squared (r2): 0.9135166954729765`
- `Mean Squared Error (MSE): 0.08139250993211926`
- `Mean Absolute Error (MAE): 0.22941606556120853`

These are slightly better parameters than those obtained previously.



	month	value	yeartodate	previousyeartodate	share	\
0	1	2.864000	84.421	NaN	0.000485	
1	1	2.864000	84.421	NaN	0.000485	
2	1	2.864000	84.421	NaN	0.000485	
3	1	2.864000	84.421	NaN	0.000485	
4	1	2.864000	84.421	NaN	0.000485	
...	...	...	...	...	...	
2257	12	65.434666	2677.672	1956.438484	0.012272	
2258	12	65.434666	2677.672	1956.438484	0.012272	
2259	12	65.434666	2677.672	1956.438484	0.012272	
2260	12	65.434666	2677.672	1956.438484	0.012272	
2261	12	65.434666	2677.672	1956.438484	0.012272	

	temperature_celsius	temperature_fahrenheit	humidity	\
0	5.5	42.0	85	
1	4.0	39.2	93	
2	10.0	50.0	58	
3	9.0	48.2	62	
4	4.0	39.2	87	
...	...	...	...	
2257	6.0	42.8	87	
2258	2.0	35.6	93	
2259	5.0	41.0	81	
2260	6.0	42.8	81	
2261	3.0	37.4	87	

	feels_like_celsius	feels_like_fahrenheit	sunrise	sunset	day
0	1.4	34.6	07:45 AM	04:10 PM	1
1	3.2	37.8	07:45 AM	04:11 PM	2
2	8.1	46.6	07:45 AM	04:12 PM	3
3	5.5	41.9	07:45 AM	04:13 PM	4
4	3.6	38.5	07:45 AM	04:15 PM	5
...	...	...	...	...	...
2257	3.7	38.6	07:45 AM	04:06 PM	27
2258	0.5	32.9	07:45 AM	04:07 PM	28
2259	2.4	36.4	07:45 AM	04:08 PM	29
2260	5.1	41.1	07:45 AM	04:09 PM	30
2261	2.0	35.5	07:45 AM	04:10 PM	31

Figure 7: IEA + World Weather Repository dataset example

7 Conclusion

This work addresses the significant issue of solar energy prediction, currently a topic of great interest in the energy industry. A considerable portion of the work focused on identifying and correcting datasets. Finding reliable datasets from official sources, containing data over a long period, proved to be a significant challenge.

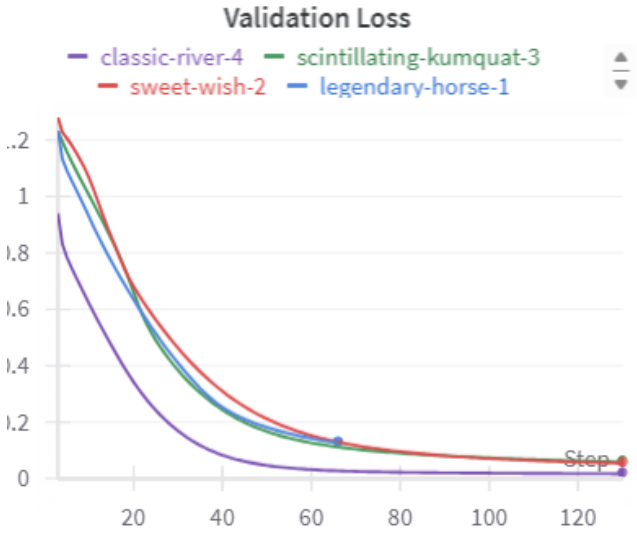
In the process of identifying datasets, crucial questions emerged: what can we do with these data? Do we need to modify some features? Are the datasets complete? Do they make sense when considered together over an extended period?

Numerous datasets did not satisfactorily answer these questions, and for this reason, they were discarded and not included in this study. The next phase of the work involved creating models based on the identified datasets. These models naturally had to make predictions. The main challenge was to identify the model that best fit the data, ensuring a good balance between performance and training time.

Further complications arose when exploring the effectiveness of combining different datasets, containing information beyond energy production, such as sunrise and sunset times and meteorological phenomena. A specific study conducted for Austria indicated a slight but significant improvement in performance.

Future developments of the project could involve the search for additional data that, combined with the already available ones, allow for more accurate predictions. Features such as cloud presence throughout the year, advanced meteorological details, and climate changes could be considered to further improve prediction accuracy.

(a) Training Curves



(b) Validation Curves

Figure 8: Curves

References

[1] Iratxe Gonzalez Aparicio et al. *Solar hourly generation time series at country, NUTS 1, NUTS 2 level and bidding zones*. [Dataset]. PID: http://data.europa.eu/89h/jrc-emhires-solar-generation-time-series. 2017.

[2] Iratxe Gonzalez-Aparicio et al. *EMHIRES dataset: wind and solar power generation*. Zenodo, 2021. DOI: 10.5281/zenodo.8340501. URL: https://doi.org/10.5281/zenodo.8340501.

[3] IEA. *IEA Monthly Electricity Statistics*. URL: https://www.iea.org/data-and-statistics/data-product/monthly-electricity-statistics.

[4] *Norway Meteorological Dataset*. Avaiable on Kaggle. URL: https://www.kaggle.com/datasets/annbengardt/noway-meteorological-data/data.

[5] *World Weather Repository*. Avaiable on Kaggle. URL: https://www.kaggle.com/datasets/nelgiriyewithana/global-weather-repository.