Small Project - EMHIRES dataset

Alessandro Castelli ID: 12246581

December 10, 2023

Contents

1 Introduction

This document aims to deal with the EMHIRES dataset, making a point of what we have done so far and trying to understand if the data we have found make sense and what we can do with this data.

2 State of the art

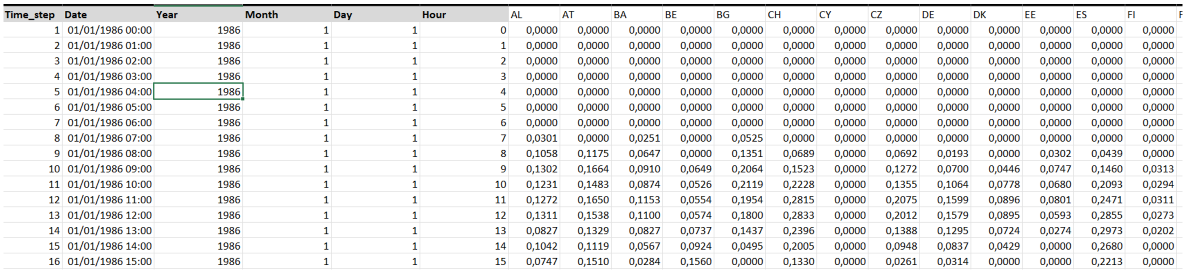
So far I have identified several datasets, one of the most relevant seemed to me EMHIRES [4]. This dataset contains European data on solar energy production available to the public derived from meteorological sources and available up to the NUTS-2 level (for more details on the dataset, see the file available on my GitHub page [2]).

2.1 Notes on the previous version

In the previous articles I worked with the dataset found on Kaggle, this dataset takes up the one released by the *European Commission* but I realized that it had some missing features that were implicit, such as the day and the hour. From now on we will consider the official dataset released by the European Commission [4].
Here are the features:

```
print(df.columns)

Output:
Index(['Time_step', 'Date', 'Year', 'Month', 'Day', 'Hour', 'AL', 'AT', 'BA',
       'BE', 'BG', 'CH', 'CY', 'CZ', 'DE', 'DK', 'EE', 'ES', 'FI', 'FR', 'EL',
       'HR', 'HU', 'IE', 'IT', 'LT', 'LU', 'LV', 'ME', 'MK', 'NL', 'NO', 'PL',
       'PT', 'RO', 'RS', 'SI', 'SK', 'SE', 'XK', 'UK'],
      dtype='object')
```

Here's an example of the first few rows of the dataset (Figure 1):



Time_step	Date	Year	Month	Day	Hour	AL	AT	BA	BE	BG	CH	CY	CZ	DE	DK	EE	ES	FI	F
1	01/01/1986 00:00	1986	1	1	0	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	
2	01/01/1986 01:00	1986	1	1	1	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	
3	01/01/1986 02:00	1986	1	1	2	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	
4	01/01/1986 03:00	1986	1	1	3	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	
5	01/01/1986 04:00	1986	1	1	4	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	
6	01/01/1986 05:00	1986	1	1	5	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	
7	01/01/1986 06:00	1986	1	1	6	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	
8	01/01/1986 07:00	1986	1	1	7	0,0301	0,0000	0,0251	0,0000	0,0525	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	
9	01/01/1986 08:00	1986	1	1	8	0,1058	0,1175	0,0647	0,0000	0,1351	0,0689	0,0000	0,0692	0,0193	0,0000	0,0302	0,0439	0,0000	
10	01/01/1986 09:00	1986	1	1	9	0,1302	0,1664	0,0910	0,0649	0,2064	0,1523	0,0000	0,1272	0,0700	0,0446	0,0747	0,1460	0,0313	
11	01/01/1986 10:00	1986	1	1	10	0,1231	0,1483	0,0874	0,0526	0,2119	0,2228	0,0000	0,1355	0,1064	0,0778	0,0680	0,2093	0,0294	
12	01/01/1986 11:00	1986	1	1	11	0,1272	0,1650	0,1153	0,0554	0,1954	0,2815	0,0000	0,2075	0,1599	0,0896	0,0801	0,2471	0,0311	
13	01/01/1986 12:00	1986	1	1	12	0,1311	0,1538	0,1100	0,0574	0,1800	0,2833	0,0000	0,2012	0,1579	0,0895	0,0593	0,2855	0,0273	
14	01/01/1986 13:00	1986	1	1	13	0,0827	0,1329	0,0827	0,0737	0,1437	0,2396	0,0000	0,1388	0,1295	0,0724	0,0274	0,2973	0,0202	
15	01/01/1986 14:00	1986	1	1	14	0,1042	0,1119	0,0567	0,0924	0,0495	0,2005	0,0000	0,0948	0,0837	0,0429	0,0000	0,2680	0,0000	
16	01/01/1986 15:00	1986	1	1	15	0,0747	0,1510	0,0284	0,1560	0,0000	0,1330	0,0000	0,0261	0,0314	0,0000	0,0000	0,2213	0,0000	

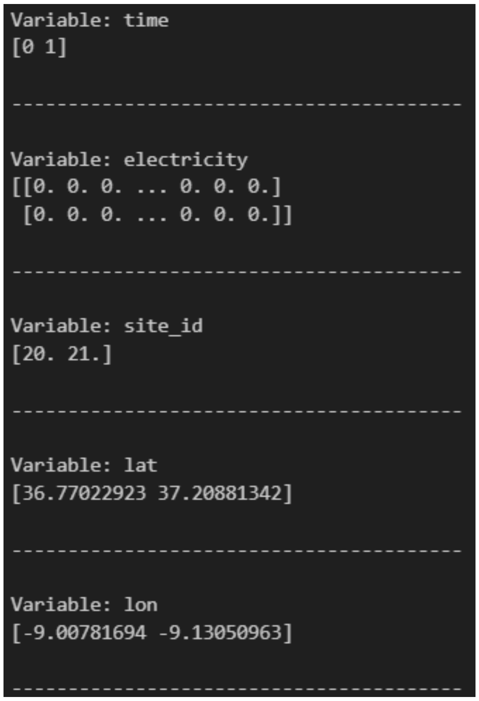
Figure 1: First rows of the dataset

3 Is this dataset valid?

It has emerged that it is necessary to identify datasets similar to `EMHIRES` to compare the data and consequently find out whether the dataset is reliable or not. Initially, I found `EMHIRES` on Kaggle [5]. Now the natural next step would be to compare this dataset with others to see if its data are consistent. Here a problem arises, I found only one dataset titled `Capacity factor time series for solar and wind power on a 50 km2 grid in Europe`[6] that seems to contain the same type of data as the `EMHIRES` dataset but there is a problem. This dataset is not a classic `.csv` dataset but is a `.nc` dataset. This type of format is mainly used when dealing with datasets with multidimensional `features` (In Figure 2 which is the output of Listing 1 you can see what the first two rows of the dataset look like, collected by feature). Also, I did not find much documentation related to this dataset and therefore it is not possible to correctly read the semantics of the features, in conclusion, the only dataset that I was able to identify for comparison cannot be compared.

```python
import netCDF4 as nc

# Specify the path of the .nc file
nc_file_path = "PATH"

# Open the NetCDF file
dataset = nc.Dataset(nc_file_path, 'r')

# Display the first few rows of all variables in the dataset
for var_name, var in dataset.variables.items():
    print(f"Variable: {var_name}")
    print(var[:2])  # Print the first 10 rows of each variable
    print("\n" + "-" * 40 + "\n")  # Add a divider line between variables

# Close the NetCDF file
dataset.close()
```

Listing 1: Extract the first two rows of the new datset



```
Variable: time
[0 1]

----------------------------------------

Variable: electricity
[[0. 0. 0. ... 0. 0. 0.]
 [0. 0. 0. ... 0. 0. 0.]]

----------------------------------------

Variable: site_id
[20. 21.]

----------------------------------------

Variable: lat
[36.77022923 37.20881342]

----------------------------------------

Variable: lon
[-9.00781694 -9.13050963]

----------------------------------------
```

Figure 2: Multidimensional features example - Feature: "Electricity"

3.1 So, is EMHIRES reliable?

There are many indications that lead us to believe that EMHIRES is a reliable dataset even though it is not possible to cross-check with other datasets.

- **Dataset Creator**: The author of the dataset is very reliable, in fact in this case, this dataset was published by the Joint Research Centre of the European Commission [3], which is an official European body. Moreover, it is the first European dataset on solar and wind energy generation publicly available up to the NUTS-2 level.
- **Data Collection Method**: This dataset was generated by applying an innovative model (the PVGIS model) that is very well done to capture local geographical information.
- **Update**: The dataset is constantly updated, the last update at the time this article is being written dates back to *September 14, 2023* [4].
- **Publications related to textitEMHIRES**: There are several publications that have been created starting from this dataset, which shows that these data have been considered reliable by the scientific community. [1], [3].

4 What Can We Do with These Data

4.1 Clustering

It would be possible to use clustering techniques to group together production sites with similar characteristics. Since *EMHIRES* is precise up to NUTS 2 level, it would be possible to accurately group which regions in Europe are more efficient in solar energy production and which regions are less efficient. To achieve this, various techniques such as **K-Means**, one of the most common clustering algorithms, could be used to divide historical data into a specified number of clusters, or through **Hierarchical Clustering**, which builds a hierarchy of clusters, allowing exploration of the structure at different levels of detail.

4.2 Prediction of Energy Production

It would be possible to train a regression model to predict future *capacity factor* projections based on the historical data studied so far. There are various methods to do this; one could use classical neural networks, or for a more advanced approach, Long Short-Term Memory (LSTM) or Gated Recurrent Unit (GRU) neural networks, designed to work with sequential data, which can be used to predict the future trends of time series.

4.3 Classification

It would also be possible to train a binary or multiclass classification model to predict whether, on a given day in a specific location, there will be *high* or *low* energy production. Naturally, to achieve this, it would be necessary to identify thresholds that define what is considered *high* versus *low*. This goal could be achieved in various ways, such as using neural networks or Support Vector Machine (SVM).

5 Conclusion

The work this week was to try to understand if the dataset is reliable, that is, if the values indicated in the dataset are actually the real values. To do this, I initially tried to identify other similar datasets with which to compare the *EMHIRES* dataset. Since I did not find datasets that contain the same types of data, I tried to orient myself on the reliability of the author and the techniques used to collect the data. It emerged that this dataset is reliable as it was created by a very important organization and that has a certain fame, namely the *European Commission*. Finally, I looked for application fields where to use this dataset, which are those explained in section 4.

References

[1] GONZALEZ APARICIO Iratxe - HULD Thomas CARERI Francesco - MONFORTI Fabio - ZUCKER Andreas. "EMHIRES dataset Part II: Solar power generation". In: (2017). URL: https://publications.jrc.ec.europa.eu/repository/bitstream/JRC106897/emhirespv_gonzalezaparicioetal2017_newtemplate_corrected_last.pdf.

[2] Alessandro Castelli. *My GitHub Page, Small Project, study EMHIRES dataset*. 2023. URL: https://github.com/Alessandro-Castelli/Small-Project/blob/main/Small_Project___Data.pdf.

[3] European Commission. *European Meteorological derived High Resolution RES generation time series for present and future scenarios.* URL: https://data.jrc.ec.europa.eu/collection/id-0055.

[4] European Commission. *Solar hourly generation time series at country, NUTS 1, NUTS 2 level and bidding zones*. 2023. URL: https://data.jrc.ec.europa.eu/dataset/jrc-emhires-solar-generation-time-series.

[5] Sohier Dane. *30 Years of European Solar Generation*. URL: https://www.kaggle.com/datasets/sohier/30-years-of-european-solar-generation.

[6] Sohier Dane. *Capacity factor time series for solar and wind power on a 50 km^2 grid in Europe*. 2022. URL: https://zenodo.org/records/6559895.