ESC 403: Group Project Proposal

Elia Torre (etorre@student.ethz.ch)
Cyrill Albrecht (cyrill.albrecht@uzh.ch)
Martha Rösler (martha.roesler@uzh.ch)
Mario Baumgartner (mariolukas.baumgartner@uzh.ch)

Solar Irradiance Prediction
31.03.2023

1.1 General topic and motivation

Many aspects of modern life are reliant on solar energy, from powering our homes and businesses to providing energy for transportation and industry. In this project, we want to accurately predict Solar Irradiance at a specific location starting from meteorological and time factors. A model of the Irradiance can help us better understand and manage this valuable resource, which in turn can lead to more efficient and sustainable energy use. This analysis has a wide range of applications, including improving solar power systems and optimizing solar energy production in different industries. For instance, precise predictions of Solar Irradiance can help solar energy firms in making informed decisions regarding the placement of new solar panels and ways to enhance their efficiency. Furthermore, this examination can also be used to anticipate fluctuations in Solar Irradiance resulting from climate change, which can assist policymakers in developing more effective measures to tackle its impact.

1.2 What data will be used?

We created five different data sets from https://power.larc.nasa.gov/data-access-viewer/ NASA platform. The POWER project combines data from different NASA research projects with the goal of an easily accessible data to study the climate. The five data sets refer to five different cities:

- Zurich [Latitude: 47.35, Longitude: 8.55]
- Milan [Latitude: 45.47, Longitude: 9.18]
- Dijon [Latitude: 47.32, Longitude: 5.06]
- Innsbruck [Latitude: 47.27, Longitude: 11.43]
- Karlsruhe [Latitude: 48.99, Longitude: 8.43]

Each data set consists of approximately 80k entries with 15 variables of which 9 hourly-measured features on a time-span of 10 years ranging from 01.01.2013 until 31.12.2022:

- The radiance [W/m^2]: The total solar irradiance incident (direct plus diffuse) on a horizontal plane at the surface of the earth under all sky conditions.
- Temperature [$C°$]: The average air (dry bulb) temperature at 2 meters above the surface of the earth in Fahrenheit.
- Humidity [g/kg]: The ratio of the mass of water vapor to the total mass of air at 2 meters.
- Precipitation [mm/h]: The total atmospheric water vapor contained in a vertical column of the atmosphere.
- Wind speed [m/s]: The average wind speed at 10 meters above the surface of the earth.
- Wind direction [°]: The average wind direction at 10 meters above the surface of the earth.
- Frost point [$C°$]: The dew/frost point temperature at 2 meters above the surface of the earth.
- Wet bulb temperature [$C°$]: The adiabatic saturation temperature which can be measured by a thermometer covered in a water-soaked cloth over which air is passed at 2 meters above the surface of the earth.
- Surface Pressure [kPa]: The average surface pressure at the surface of the earth.

And additional variables such as:

- Timestamps: year, month, day, hour.
- Time of sunrise.
- Time of sunset.

Why are we expecting a relation between solar irradiation and those factors?
From a model perspective, the sun has a constant radiation. On a point on the earth the incoming radiation fluctuates due to the night-day cycle (i.e., time) and decreases also through absorption and scattering in the atmosphere. The amount of this decrease in the radiation is dependent on the medium it travels through, therefore it makes sense to discuss the effect of different states of the atmosphere.

1.3 What questions will be asked and attempted to answer?

Our main focus will be on performing an in-depth statistical analysis of the Zurich data set and the relative solar radiation prediction task. Specifically we want to answer the following questions:

- What is the relationship between solar radiation and local climate patterns, such as temperature, humidity and precipitation?
  - Are there non-trivial relations between the different variables?
- Can we find an accurate machine learning model to predict solar radiation based on the meteorological variables and historical data?
  - Which variables have the biggest impact and which ones are negligible?
  - Which model best explains the relationships among variables? Why?
  - Does the prediction reach a level of accuracy such that it could be exploited for application purposes?
- In addition, we chose other 4 locations: Dijon, Innsbruck, Karlsruhe and Milan. [Dijon, Innsbruck, and Zurich] share nearly the same Latitude. While [Karlsruhe, Milan, and Zurich] share nearly the same Longitude. This results in a "cross" with Zurich in the middle. We are wondering whether it is possible to predict the solar radiation in Zurich as an average of the radiations in these 4 other locations.
  - Suppose we are a company interested in estimating the solar radiation in a location of which we do not have a direct measurement. Is it possible to estimate this measure by averaging over 4 other neighboring locations? What gives us the most reliable estimate?
    - Averaging the measured solar radiation in the 4 cities?
    - Training an model on a custom data set featuring information of all of the 4 cities?
    - Ensembling the measured solar radiation in the 4 cities with different weights according to a similarity measure between Zurich climate and the other cities climates?
  - Which of the 4 neighboring locations has the most similar climate to Zurich? Does it also share the highest correlation in terms of solar radiation over time?
- We would have been also interested in evaluating the role played by air pollution in diminishing the amount of solar radiation that reaches the earth surface, as well as, how that affects the other meteorological factors (e.g., temperature). However, we were not able (after an extensive search) to identify a reliable data source providing information regarding air pollution with the desired time span and resolution in the 5 locations designated for this analysis. We are open to suggestions if you are aware of any source providing this data.

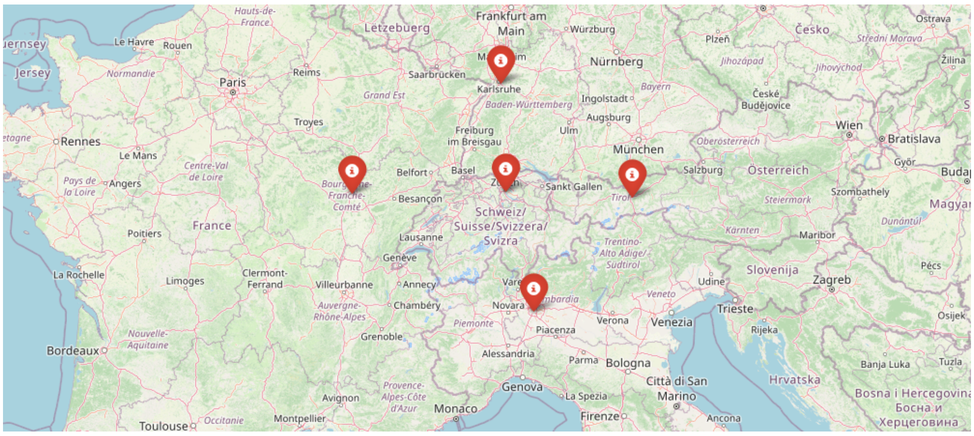


Figure 1: Plot with Folium library of the 5 locations

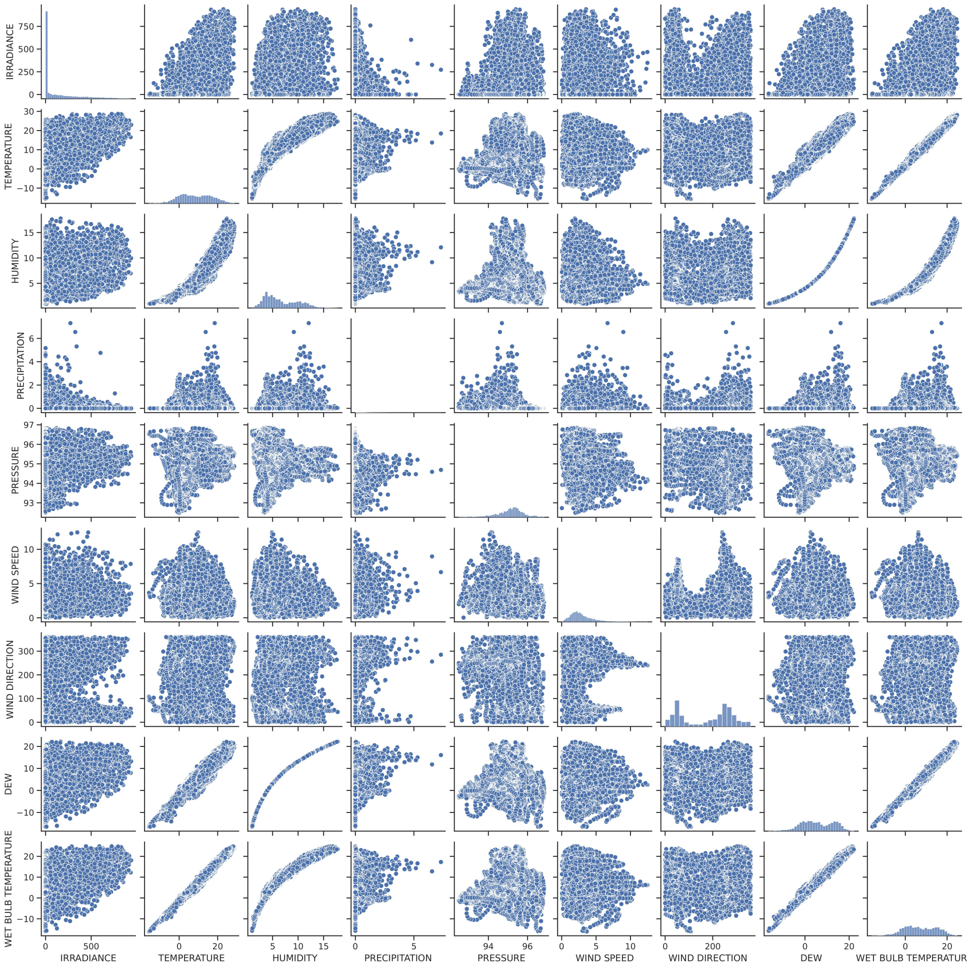
1.4 How will the data be processed?

Firstly, we perform some *data exploration* to better grasp the data at hand. We will then proceed in a *data cleaning* process (missing values, outliers, incorrect data, anomaly detection), followed by *data augmentation* (we added the time of sunrise and time of sunset via Astral library). After this process, we will perform *descriptive analyses* of the data set investigating variables relations/correlations, as well as possible trends. We will then investigate possible *feature engineering* work to be done, especially in regards to the treatment of time variables to be able to exploit them in the prediction task (transforming the time of sunrise/sunset into "how much time elapsed after sunrise"/"how much time till sunset"). Then, after a thorough *exploratory* and *descriptive analysis* of the data set we will proceed to the prediction task by performing *model selection*, *tuning*, *cross-validation* and *ensembles*. Finally, we will draw our conclusion regarding whether it is a feasible prediction and assess whether the "quadrangulation" over the 4 neighboring cities gives us reliable estimates of radiation in Zurich.

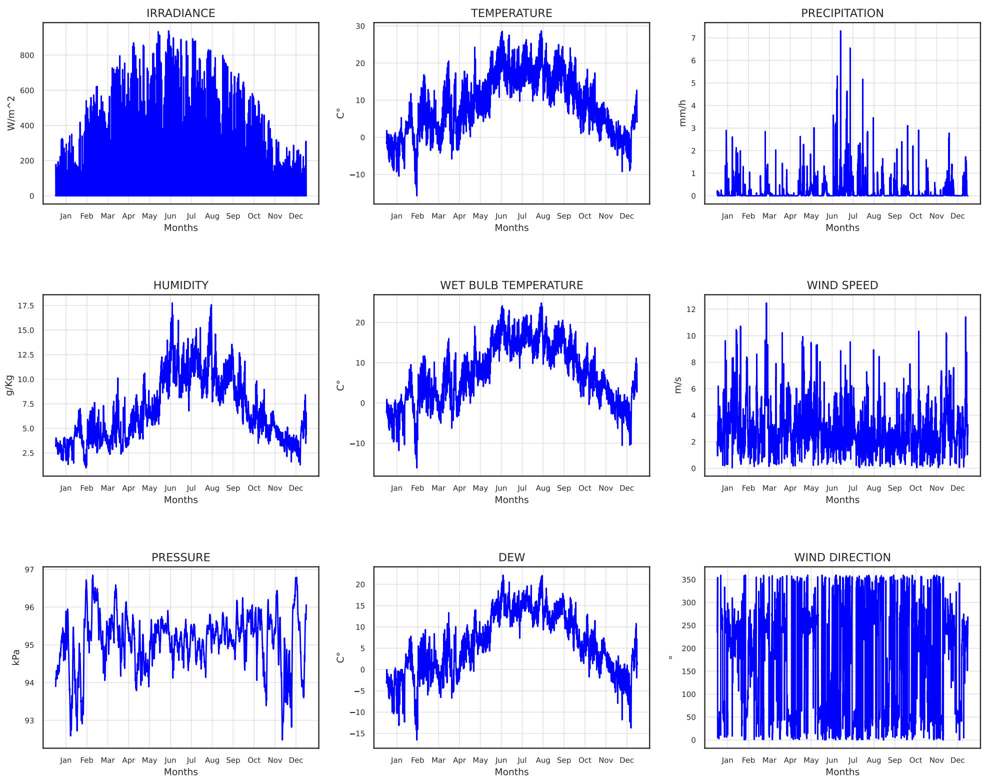
1.5 What analysis techniques or algorithms are planned to be used?

- **Data Exploration & Description**: EDA, Variables Distributions (Histograms & Plots), Pairwise relationships, Correlation Heat Map.
- **Data Cleaning**: Missing Values and Outliers Detection.
- **Data Augmentation & Feature Engineering**: Astral library for Time Of Sunrise and Time of Sunset, Transformation in "how much time elapsed after sunrise" and "how much time till sunset", (possible) standardization and/or transformation of the features for better fitting.
- **Model Selection**: Linear Regression, K-Nearest Neighbor Regression (KNN), Decision Tree & Random Forest, Light Gradient Boosting Machine (LightGBM) & Extreme Gradient Boosting (XGB), AdaBoost/CatBoost, Feed-Forward Neural Network (FFNN), Recurrent Neural Network (RNN), Long-Short Term Memory RNN (LSTMs), we will look at the performances obtained by the various architectures and understand where to orient ourselves.
- **Tuning & Ensembles**: Randomized and/or Grid Search for optimal model hyperparameters, Averaging Algorithms, Voting & Stacking Regressors.
- **Cross-Validation & Evaluation**: KFold Cross-validation, Mean Squared Error (MSE), Root Mean Squared Error (RMSE), R-Squared (R^2).

1.6 Plot of the pairwise relationships of the data set of Zurich from 2021



1.7 Plot for each feature of the data set of Zurich from 2021



As one can see by looking at the plot it's already possible to observer some correlations between values of the data set and the irradiance.

1.8 Correlation-matrix of the data set of Zurich from 2021