**Assignment I.**

**Analysis of photovoltaic (PV) power plant located near Tallinn.**

**Ilja Matjas**

**Taltech / Tallinn 2024**

[**ilja.matjas@taltech.ee**](mailto:ilja.matjas@taltech.ee)

**Introduction**

Home assignment (hereinafter HA) is based on analysis of photovoltaic (hereinafter PV) power plant (hereinafter PP) located near Tallinn. Input data and Scope of Work (hereinafter SoW) were provided by course instructors.

Provided data is time-series and includes hourly electricity production data and temperature data.

Multiple analytical and coding tasks, as defined in SoW, are done within HA. Result, finding and discussions are presented in this report.

**Business understanding**

Analysis and modelling PV production can be useful at least in terms of commercially valuable insights and scientific research (R&D). For sure, there are still dozen more useful stuff to be done with such data, but I will stick to these two.

* Commercial value:
  + accurately predicting production of PV for next hour, day, whatever granularity might be needed, to minimise production-consumption balancing costs
  + accurately predicting production of PV to make correct decisions on energy futures markets
  + further investment decisions based on performed analysis (shall I invest in one more PV PP)?
* Scientific research:
  + insights for planning energy power system adequacy (adequacy means that power system generation is always possible to cover power system load)
  + insights for research and modelling if and how to use PV PPs in frequency supporting services (FCR, FRR). PV PP is of-course not 100% reliable energy source, but maybe there is some way to use them as reserves based on probabilistic models.
  + use such data as a measurement tool to track trends related with nature phenomenon. Maybe can do some conclusions or even use as a part of some nature-descriptive model if production is rising/decreasing as a trend from year-to-year for example?

**Data collection**

**Understanding available data**

Data is provided in tabular form and contains 17016 rows and 6 columns. Columns included in dataset are represented in Table XXX

|  |  |  |  |
| --- | --- | --- | --- |
| Column name | Value example | Count | Data type |
| timestamp | 1 612 224 000 000 | 17 016 | int |
| raw | 178 582.47 | 17 016 | float |
| temperature | -3.9 | 17 016 | float |
| modified1 | 0.60 | 15 825 | float |
| modified2 | 172 644.91 | 15 641 | float |
| modified3 | 17.34 | 15 879 | float |

Table XXX

From SoW, Table XXX and data itself we can understand what kind of data is provided. Timestamp is a unix-timestamp given in miliseconds. We can understand that these are miliseconds from power of values e.g. 1.6e+12. In seconds the order would be e.g. 1.6e+9. Time step is 1 hour or 3 600 seconds, however it is not always the case in provided data. Different time steps and their counts are given in Table XXX.

|  |  |  |
| --- | --- | --- |
| Time step (s) | Time step (h) | Count |
| 3 600 | 1 | 17 004 |
| 90 000 | 25 | 10 |
| 954 000 | 265 | 1 |

Table XXX

From table XXX we can see that for ten times data is missing completelly daily. And once is missing for 11 days. Some hypothesis might be raised why this data is missing. For example, daily outages could be maintanance works and longer 11 day period might reflect unplanned breakdown outage.

Raw is an hourly production of PV PP. As maximum value of raw equals to 301 630 we can with very high confidence conclude that measuring unit is Wh (Watt-hour). Raw can not be given in kWh, bacause physically there is no 300 MW large PV PP installed in Estonia. However, it is more convinient to work in kWh scale, so I converted raw column from Wh to kWh.

Temperature is given in Celsius. This can be concluded from derivng minimum and maximum values from the column. Minimum value is -24.5 and maximum 30.1. Thus, we can be confident that data is not provided in Farhengeit or Kelvin.

Modified1, modified2 and modified3 columns are randomly or using some logic, modified raw column. Thus, these values are also provided in Wh.

**External data sources**

It depends on scope of analysis and modelling is any additional sources might be needed.

To understand what the provided data is, what are the measuring units, etc. I didn’t need to use additional external resources.

To catch more patterns or to end up with nice predictive model some additional sources shall be very useful. For example, cloudiness has very large effect of PV PP production due to reduced level of sun insolation. Also wind strength, as wind is acting as a cooling, it allows to generate more.

As we are dealing with PV PP, sun map and sun-clock information might be very useful.

Information related to planned (and maybe unplanned) outages of PV PP might be very useful, specially if outages are in correlation with weather conditions or happen periodically. Electricity price data might be useful, as plant could be just turned off if the price went negative (cheaper to shutdown the plant instead of paying for production).

To complete HA, I haven’t used any additional external data sources.

**Data cleaning**

After reading the data, missing values check was done. Results of that check are represented in Table XXX.

|  |  |  |
| --- | --- | --- |
| Column | Count of missing values | % of missing values |
| Timestamp | 0 | 0 |
| Raw | 0 | 0 |
| Temperature | 0 | 0 |
| modified1 | 1 191 | 7.00 |
| modified2 | 1 375 | 8.08 |
| modified3 | 1 137 | 6.68 |

Table XXX

As it can be seen from the Table XXX only columns modified1, modified2 and modified3 contains missing values.

To get understanding about the nature of missing data some steps were taken. First of all, I binary visualised the patter of missing data. This binary sequential representation is given on Chart XXX.



Chart XXX

White strip on black font is representing occurrence of missing value. From Chart XXX one can see that nature of missing value of modified3 is completely different compared to modified1 and modified2. However, we need more information to conclude mechanism of missing data. To collect more information, a correlation analysis was done.

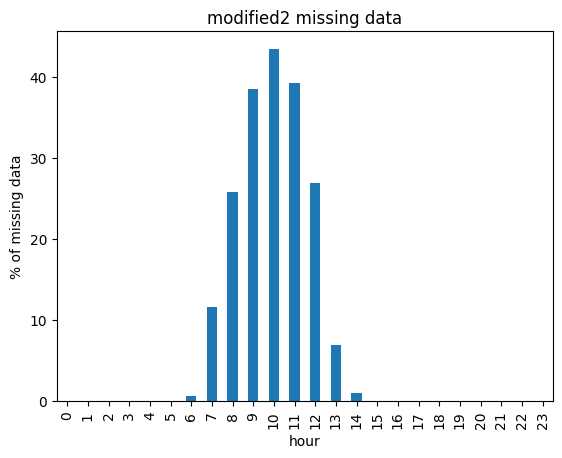
For correlation analysis I used temperature as a feature and additionally generated three more time-based features: day of week, month and hour. To catch correlations between missing values of modified columns and features mentioned above, modified columns were binary encoded such, that values with present data were set to 1 and with missing values were set to 0. For day of week, month and hour features one-hot-encoding was applied, setting values to 0 and 1 as well. Top 6 correlations for each modified feature are represented in Table XXX

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **modified1 correlations** | | **modified2 correlations** | | **modified3 correlations** | |
| day\_of\_week\_3 | 0.026 | hour\_10 | -0.271 | temperature | -0.501 |
| hour\_6 | 0.017 | hour\_11 | -0.238 | month\_6 | -0.197 |
| hour\_7 | 0.017 | hour\_9 | -0.233 | month\_7 | -0.145 |
| hour\_23 | 0.016 | hour\_12 | -0.144 | month\_8 | -0.109 |
| day\_of\_week\_1 | -0.015 | hour\_8 | -0.136 | month\_9 | 0.06 |
| hour\_16 | -0.013 | temperature | -0.116 | month\_3 | 0.06 |

Table XXX

From Table XXX it is seen that missing values of modified1 are not in correlation with any other feature. Thus, it can be concluded that modified1 is missing completely at random or (**MCAR**).

Missing values of modified2 is strongly correlating with certain hours (8-12) with peak at hour 10 and is correlating with temperature. As data is never missing during night or twilight hours, and distribution of missing data is very similar to distribution of mean hourly production, it can be concluded that there is a relationship between them. So, it can be related to measurement malfunctions due to sun insolation intensity. Such missing data mechanism can be concluded as missing at random or (**MAR**).

 A graph of a number of hours

Description automatically generated

Chart XXX Chart XXX

Missing values of modified3 is very strongly correlating with temperature (and high temperature obviously is in high correlation with summer months). Chart XXX shows distribution of missing data with respect to temperature. At high temperatures data is not being measured or transmitted at all. This can be since with high temperatures, generation and production of PV PP is so high, that measurement system cannot handle such high datapoints and thus they are not being transmitted. If so, then the reason of missing data is the data point itself. Such missing data mechanism can be concluded as missing not at random or (**MNAR**).

A screenshot of a computer

Description automatically generated

Chart XXX

**Dealing with missing data**

First and very simple way to deal with missing data is just to drop it. As already stated in Table XXX, after dropping missing data of modified1, dataset shall loose 7% of data.

More advanced way to inpute missing values would be to build a model which then would fill missing values. To make it accurate, nature of the data and mechanism of missing data shall be taken into account.

We already concluded that missing values are MCAR, so no sophisticated modelling approach shall be required.

We also know that we are dealing with sequential data. In sequential data, points located closer to each other, often carry the most information regarding neighbouring points. Thus, it would be a good idea to do single imputation grouping data daily e.g. (to fill ’2022-01-01’ missing values, ’2022-01-01’ hourly data shall be used) and for multivariate imputation averaging same hours of neighbouring days e.g. (to fill ’2022-01-01’ missing values, ’2021-12-31’, ’2022-01-01’ and ’2022-01-02’ mean hourly data shall be used)

From Chart XXX, based on modified1, we can see mean profile of daily PV PP production. It is quite accurately bell-shaped. Even-order polynomial would be a good choice to fit such data. To mimic bell-shape 6’th order polynomial was used and result is also given on Chart XXX.

Polynomial will also return negative numbers. In our case, negative numbers can be converted to zeros, as we have the knowledgebase that during whole period of measurements PV PP was not acting like a consumer, thus production shall not go to zero.

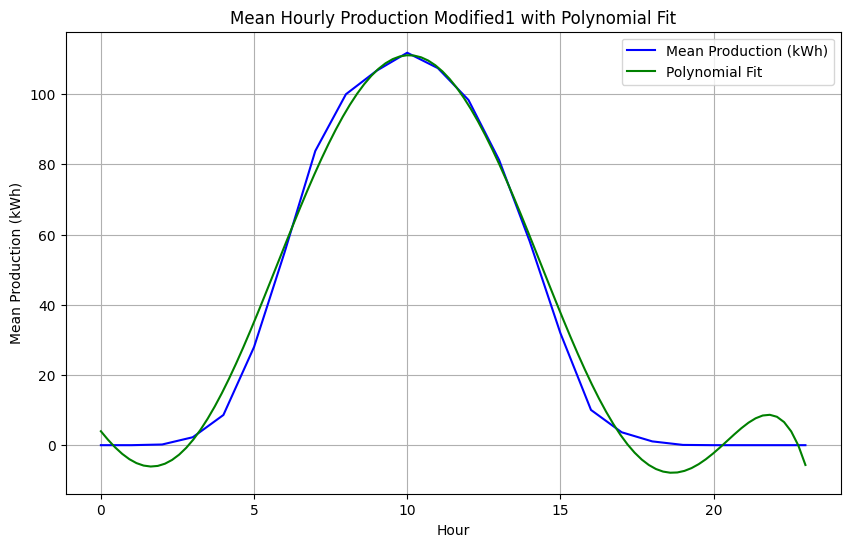


Chart XXX

Performance metrics of different imputation metrics, based on comparison between raw and modified1 data is represented in Table XXX

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Imputation Method | Negative value shaving | MSE / SE | MAE / SE | MAPE / SE |
| Single | No | 78.30 / 16.50 | 1.19 / 0.067 | 5.93 / 3.76 |
| Yes | 35.94 / 4.0 | 0.81 / 0.046 | 2.46 / 1.64 |
| Multivariate | No | 62.37 / 5.65 | 1.17 / 0.06 | 3.26 / 1.91 |
| Yes | 60.61 / 5.65 | 1.04 / 0.059 | 1.67 / 1.15 |

Table XXX

MAPE metric used in the project is quasi-MAPE as all zeros from actual values were discarded to calculate it (otherwise MAPE would have been infinity). From MAE and MSE we can see that multivariate imputation performed better compared to single imputation. It is because model can generalise better as number of instances to build the model was higher. Zeros shaving also helped, especially in scenario with single imputation. Multivariate model is more robust towards not going to negative part as it had more data to generalise over hours. Large values of MSE are signalising that some missing points were badly predicted. One of examples of badly predicted missing value is shown on Chart XXX. Overall, results of imputation can be concluded as satisfying.

A graph of a line graph

Description automatically generated with medium confidence

Chart XXX

**Data exploration and Feature engineering**

Statistical description of raw (production) data is given in Table XXX. From statistical description we can see that standard deviation is almost twice as high as mean. It indicates on large variability in the data such as high spread, possible outliers or skewness. If we look further, we can also notice that median is zero. It means that at least 50% of datapoints are equal to zero. So, data is having right-skewed distribution. Maximum value is almost 8 times larger and 90% quantile almost 5 times than mean. Thus, there is a spread in the data.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Metric** | Count | Mean | Std | Min | Max | Q 1/4 | Q 1/2 | Q ¾ | Q 9/10 |
| **Value** | 17 016 | 36.91 | 66.91 | 0 | 301.63 | 0 | 0 | 38.26 | 154.99 |

Tabel XXX

Distribution represented on Chart XXX confirms our conclusion above. Data is indeed right skewed with significant spread and zeros occurring in more than half points.

A graph of a distribution of production data

Description automatically generated

Chart XXX

Such zero-rich skewness can be considered as extremely skewed data. I am not sure if there is a way to apply a transformation on such data to get data close to normally distributed and avoid data loss (e.g. be able to represent 100% of data via back-transform).

However, we have a knowledge base (and we could also do quick analysis to verify it) that during no-sun hours there is no PV PP production. So, first simple and robust idea to apply as data transformation would be to drop night hours. For that additional external data or third-party library could be used. Or we can derive zero monthly hours from data. From charts XXX and XXX one can see how zeros are related to hours and months. After dropping non-sunny hours (search was done within data itself) distribution became better (Chart XXX), but still contains quite many zeros. So, to be able to transform data as normal distribution I had to discard zero values. After that, applying nested transformations (power 1/3 -> power ½ -> log\_e (3 – data) I was able to come up with distribution represented on Chart XXX. I also tested back transformation and managed to get original data. Thus, transformation did not result in any data loss.

A graph of a number of bars

Description automatically generated with medium confidence

Chart XXX

A screen shot of a graph

Description automatically generated

Chart XXX

A graph of a number of production data

Description automatically generated

A graph of a distribution of production data

Description automatically generated

**Additive classical decomposition**

In PV PP data trend could have been derived from yearly productions, because quarterly and monthly data has very strong seasonal component (see Chart XXX). However, data contains only two years, so we have 2 datapoints to model trend. We’ve could make a linear approximation of the trend between these two points, but that would be rather approximation than a comprehensive trend over a longer period. Thus, in terms of this HA I neglect trend component and set it to zero T(t) = 0.

A graph of a line graph

Description automatically generated with medium confidence

Chart XXX

To calculate seasonal component S(t), points representing each season were averaged. As I concluded that trend is neglected, no de-trading prior calculating seasonal component was needed. Results of S(t) are given on Chart XXX.

Finally, residual component can be calculated by subtracting both trend T(t) and seasonal S(t) components. Result of R(t) is given on Chart XXX

A graph with numbers and lines

Description automatically generated

A graph with blue lines

Description automatically generated