Spatial Data Analysis and Modelling on Norwegian Windfarm Data

PROJECT REPORT

In the **MOD550** course of the master program **Computational Engineering** at Universitetet i Stavanger

by

René König

Atanu Das

29th May 2022

Student numbers	268127, 268128
Course	MOD550
Supervisors	Prof. Reidar B Bratvold Muzammil H Rammay Aojie Hong

Abstract

In this project spatial data analysis and spatial modelling is conducted on Norwegian windfarm data as well as wind resources data to gain insight in factors determining suitable locations for new windfarms. Data from the Norwegian Water Resources and Energy Directorate on average windspeed and terrain complexity is mapped together with data on licensed power and expected energy production of existing windfarms and subsequently sampled at their location to be analysed together using scatterplots and histograms, showing that windfarms are primarily placed in areas with low terrain complexity. Additionally, the effect of the construction date of the windfarms on their licensed power and expected energy production is examined, confirming that newer windfarms are more efficient and powerful than older windfarms. The spatial modelling techniques de-clustering, experimental variograms, variogram modelling and ordinary kriging are applied on the estimated energy production data of the windfarms resulting in a spatially interpolated kriging map.

Table of contents

List of figures

List of abbreviations

NVE	Norwegian Water Resources and Energy Directorate

1 Introduction

The purpose of this project is to apply the knowledge gained from the course "MOD550: Applied Data Analytics for Spatial and Temporal Modelling" on a chosen problem. The methods have primarily been discussed in the course in their applications on problems from the oil- and gas sector. Since these methods should be universally applicable, we want to apply some of those methods, such as de-clustering, variogram modelling and kriging, on data from the Renewable Energy sector as this sector is gaining more importance in the recent years.

Because of environmental as well as political concerns, the energy sector is more and more shifting towards renewable energy sources like solar, water and wind energy. While these resources are renewable, the locations at which they can be harnessed are limited by environmental factors as well as needing to be located close to the places where the energy is needed, since electrical energy transport over long distances is inefficient and costly. Therefore, the locations for new renewable energy projects have to be chosen carefully.

We chose to work on data from Norwegian windfarms and wind resources, conducting spatial data analysis and applying spatial modelling techniques such as de-clustering, variogram modelling and kriging to gain insight on factors determining good locations for new windfarms.

2 Methodology

Various data on windfarms and wind resources from the Norwegian Water Resources and Energy Directorate (NVE)´s Map services has been examined.

Spatial data analysis is conducted on three selected datasets for whole Norway. Afterwards spatial modelling is conducted on a reduced version of the primary data covering south-western Norway. The methodology in spatial data analysis and spatial modelling is described below. Some of the results are presented and discussed in the section “3 Discussion and Results”.

2.1 Spatial Data Analysis

The primary data used is the data on Windfarms, including Windfarms under construction as well as operational ones. It is provided as an excel file and includes, among other data, the X- and Y- coordinates of windfarms, their licensed power in MW and expected energy production in GWh. The original dataframe consists of 263 datapoints.

The secondary data used is the data on yearly averaged windspeeds at 50m above ground. It is provided as an ASCII file and includes the average windspeed for each grid cell defined by columns separated by whitespace and rows separated by new lines with a gridsize of 1000m x 1000m. The windspeed is given as integer in m/s, missing values are represented by “-1”. The original dataframe consists of 3,560,112 datapoints.

The tertiary data used is the data on terrain complexity. It is provided as an ASCII file of the same structure as the windspeed data and a grid size of 200m x 200m. The original dataframe consists of 46,709,120 datapoints.

2.1.1 Data Import and Matching

The windfarm data is imported using the function “read_excel” from the “pandas” library, using only the relevant columns for X- and Y- coordinate, licensed power, expected production, number of turbines (not used) and construction date. Datapoints with nulls in the licensed power and expected production as well as outliers with an expected production above 3000 GWh, which are mainly offshore windfarms, are dropped. The indices of the dataframe are reset. The final dataframe used for the spatial data analysis consists of 256 datapoints.

The wind and terrain data are imported using the “read_csv” function with whitespace as the separator. They are scaled and matched up to the windfarm data using the provided meta-data on the coordinates of the lower left corner, cell size, number of cells in x- and y- direction and an offset in y-direction that had to be manually found through a trial-and-error process and applied to the wind data to match it up with the windfarm data.

The expected windfarm production data is visualized over the maps of average windspeed and terrain complexity.

Because information on the wind direction was not available from the processed data, it has been taken from an external source. Figure 1 shows the primary wind directions in Norway, where the distance from the centre indicates the hours per year that the wind blows from the indicated direction and the windspeed is indicated by the colour.

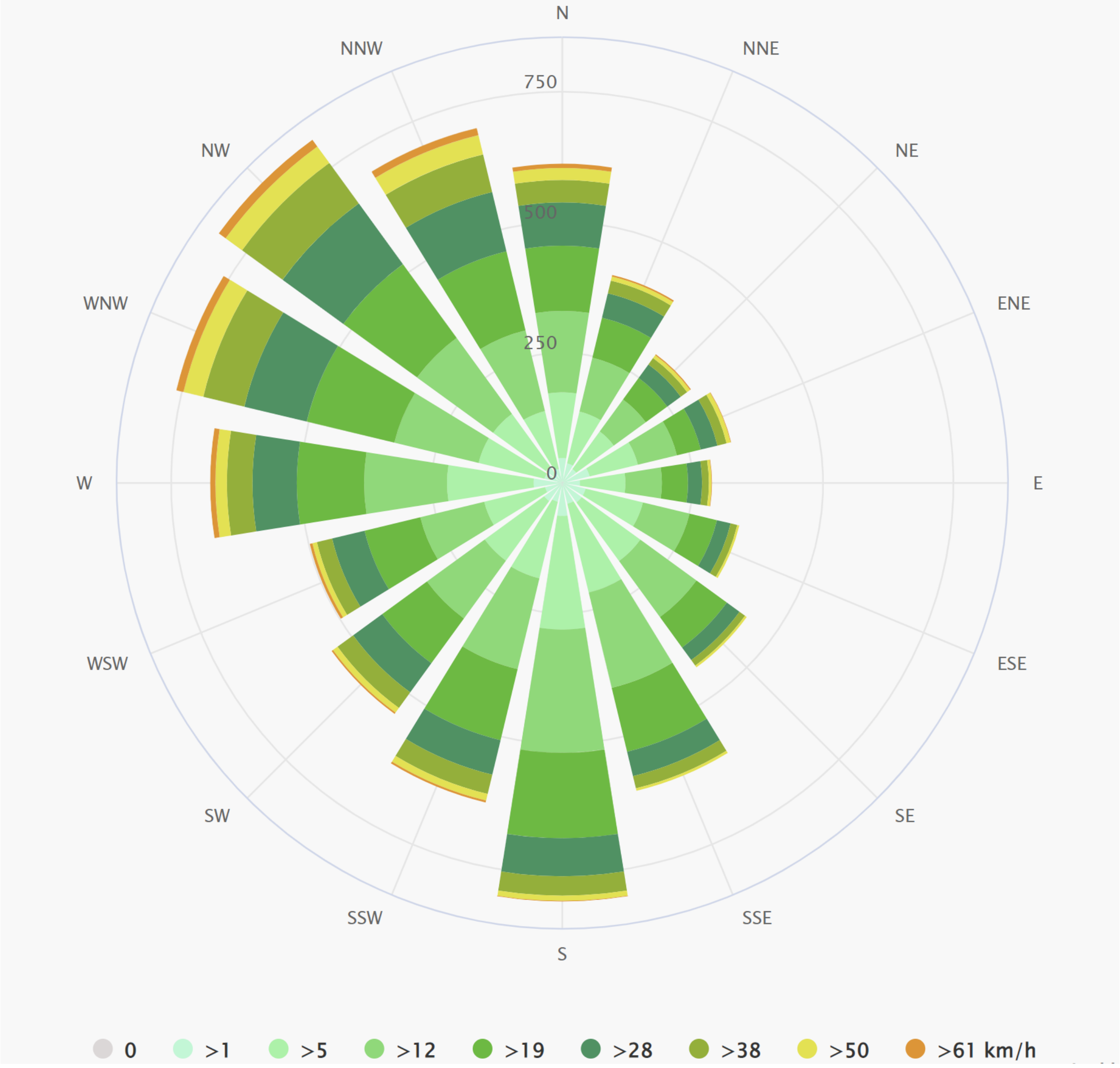

Figure 1: Primary Wind Directions in Norway [1]

From Figure 1 we can see that the wind is primarily coming from north-westerly and southern directions. This has been taken into consideration during the data analysis.

2.1.2 Sampling at Windfarm Locations

The wind and terrain data are then sampled at the locations of the windfarms by finding the respective cell with its cell centre closest to the coordinates of the windfarm and adding the cells value to the windfarm datapoint in the windfarm dataframe.

2.1.3 Plots

Spatial data analysis is conducted in form of automated plots, histograms, scatterplots, scatterplots with color-coding and 2D histograms using for-loops over the pre-selected columns of terrain complexity, average windspeed, licensed power and expected production in the windfarm dataframe.

Covariance and correlation matrices are calculated and a closer look is taken at the construction date, where that data is available.

2.2 Spatial Modelling

De-Clustering-, experimental-semi-variogram, variogram-modelling- and Kriging techniques are applied on the windfarm data for south-western Norway to create a map of optimal windfarm locations.

2.2.1 Data Import

The windfarm data is re-imported for the spatial modelling to start with a clean dataframe, only taking the columns for X- and Y- coordinates and expected production.

Nulls and outliers are removed as before and all datapoints with X- coordinate higher than 400,000 and/or Y- coordinate higher than 7,200,000 are dropped to limit the dataset to the south-western region of Norway, where most of the windfarms are located, due to restrains on computational resources.

The indices of the dataframe are reset. The final dataframe used for the spatial modelling consists of 189 datapoints.

2.2.2 Distance and Angle Calculation

As preparation for the following steps the X- and Y- distances from each datapoint to each other datapoint, as well as the Euclidian distance, angle and heading towards each other datapoint are calculated and added to the dataframe

2.2.3 De-Clustering

De-Clustering is performed on the datapoints to remove sampling bias by first defining a de-clustering grid with a grid origin and grid size. Different grid sizes and origins have been tested and a grid size of 10 x 10 cells with an origin 1 unit west of the most western datapoint and 1 unit south of the most southern datapoint are chosen.

For each datapoint the index of the corresponding grid cell is written into the dataframe. From the dataframe the number of datapoints per cell and the number of occupied cells is calculated. Using this information for each cell the cell weight is calculated and applied to the datapoints in that cell.

Figure 2 shows the windfarm locations in south-western Norway with the de-clustering grid and the de-clustering weights.

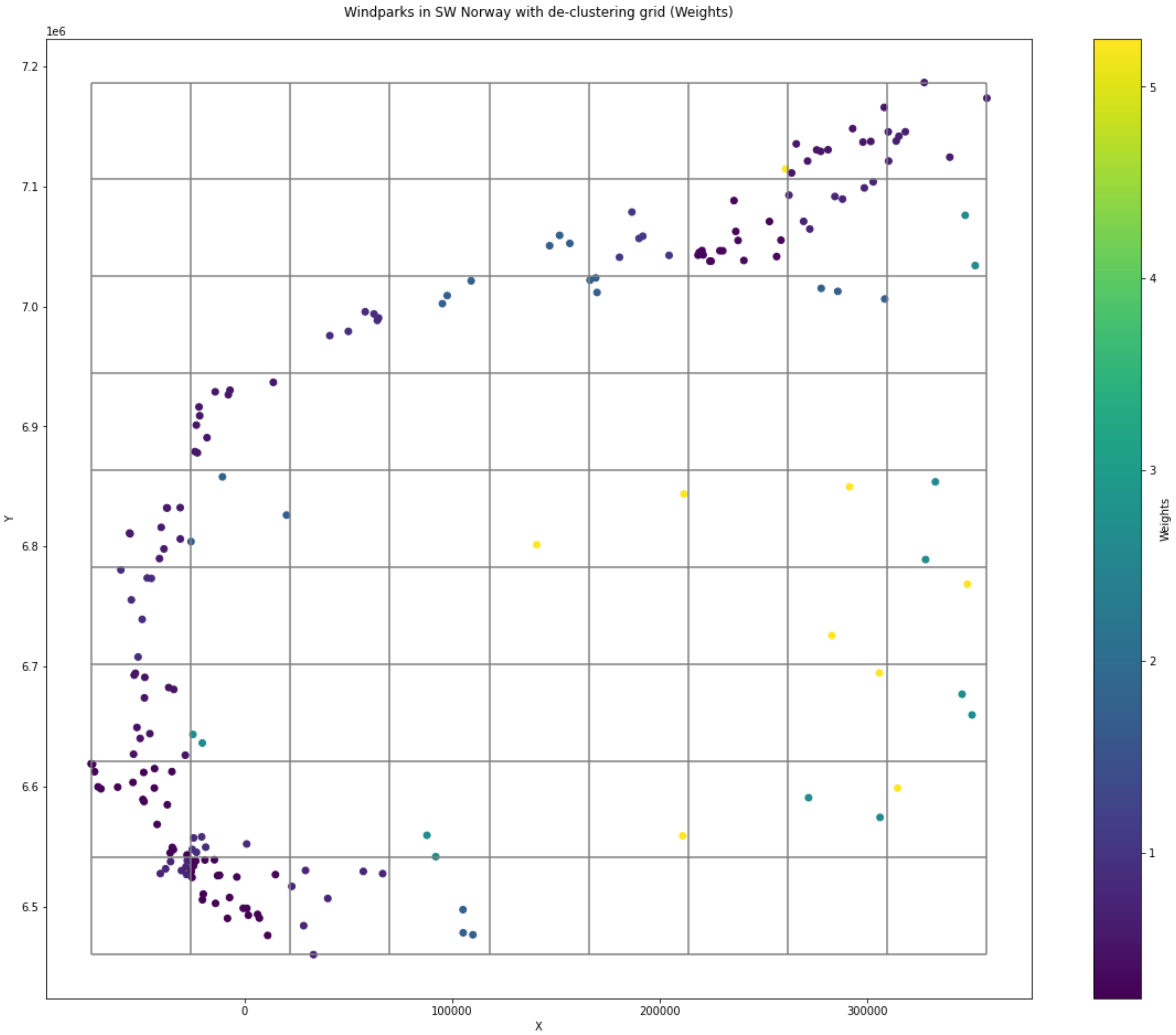

Figure 2: Map of windfarm locations with de-clustering grid and weights

From Figure 2 it is visible that cells with a low number of datapoints are assigned a high weight while cells with a higher number of datapoints are assigned a low weight. The naive data and weighted data are visualized on the datamap through color-coding and displayed together with the de-clustering grid.

2.2.4 Experimental 2D Semi-Variogram

An experimental 2D semi-variogram is created using the de-clustered data as a basis for the variogram modelling step, which is needed for the Kriging process.

A function is defined to check, if a given datapoint is inside a given lag of another given datapoint, returning True or False. Another function is defined to calculate gamma based on the value of the origin node and the values of other nodes.

The direction vector, angle tolerance, bandwidth, lag distance, lag tolerance and number of lags are specified for the experimental 2D semi-variogram and after a trial-and-error process have been chosen as shown in Table 1:

Table 1: Experimental 2D Semi-Variogram Parameters

Parameter	Value
Direction Vector	42°
Angle Tolerance	18°
Bandwidth	60,000m
Lag distance	50,000m
Lag tolerance	25,000m
Number of Lags	12

Using the previously defined check-in-lag-function, a 3D-boolean-array is created for each lag - origin-datapoint - datapoint combination. This array indicates if a given datapoint is in the given lag of a given origin-datapoint.

From this Boolean array the indices of relevant datapoints for each lag - origin-datapoint combination are taken to calculate the gamma values for all lag – origin-datapoint combinations. For that the previously defined gamma function is used, using the weighted values from the de-clustering step for the origin-datapoints and the corresponding relevant datapoints for the lag – origin-datapoint combination.

Figure 3 shows the relevant datapoints in lag 2 of origin-datapoint 12 in orange as well as the direction vector (red line) and the angle tolerance (blue lines). Note that the parts of the lines extending behind the origin-datapoint should be ignored. Also note, that the limits due to the bandwidth tolerance are not shown and the lag tolerance lines have been added in postprocessing the better visualize the lag. Similar figures to Figure 3 have been used to examine different origin-datapoint – lag combinations by changing the Point-of-interest (POI) and Lag-of-interest (LOI) values in the code for the figure in the trial-and-error process for finding suitable parameters for the experimental 2D semi-variogram.

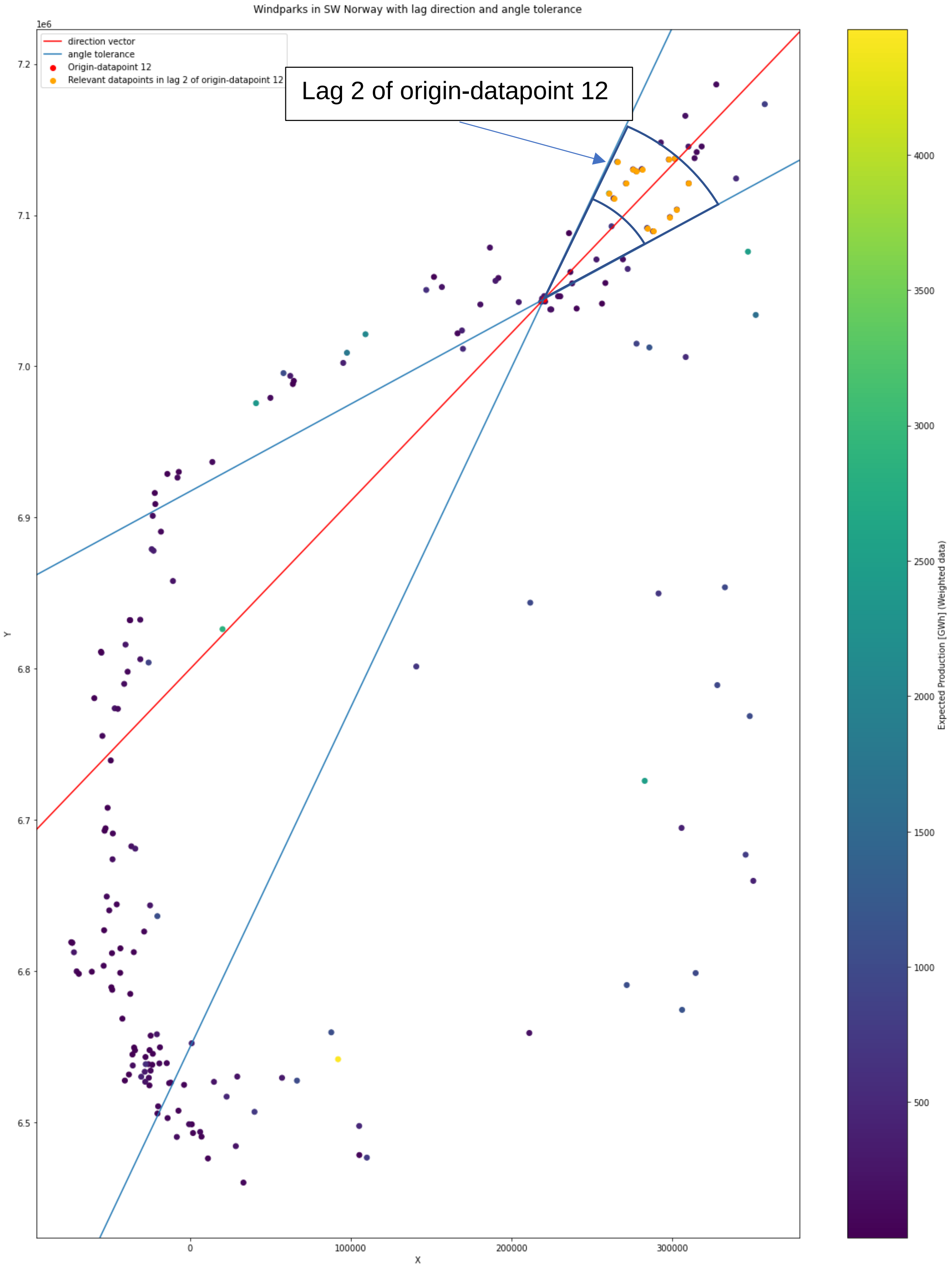

Figure 3: Visualisation of experimental 2D semi-variogram parameters

Figure 4 shows the corresponding semi-variogram for origin-datapoint 12.

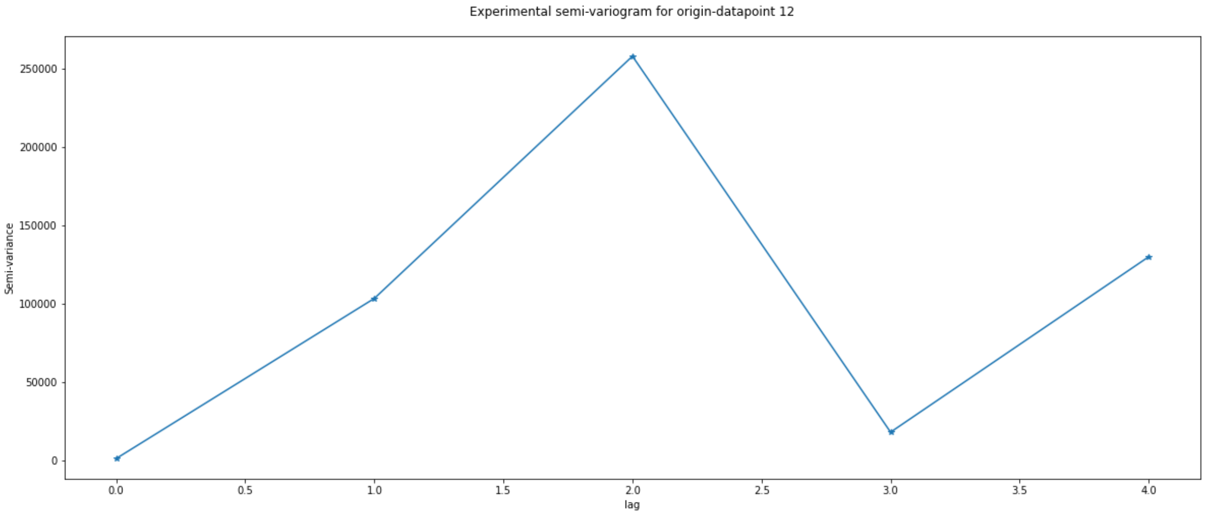

Figure 4: Experimental semi-variogram for origin-datapoint 12

The average of the semi-variograms of all origin-datapoints has been taken to create the final experimental semi-variogram. Note that at some lag – origin-datapoint combinations there might be null values, since there simply might not be any datapoints in that lag, as for example in all the lags bigger than 4 for datapoint 12, as can be seen from Figure 4. Therefore null values need to be ignored in the averaging process to get a consistent variogram.

2.2.5 Variogram Model Fitting

Functions for the spherical, exponential and gaussian variogram model are defined. The “curvefit” function from the “scipy.optimize” library is used to fit these models to the datapoints of the averaged experimental semi-variogram.

Lower and upper bounds, based on the data variance and what can be seen from looking at the experimental variogram, must be defined for the range, sill and nugget to achieve a good variogram model that represents the overall behaviour of the data, especially at close range and ignores the variation at longer ranges.

Through a trial-and-error process of different variogram models in combination with different experimental semi-variogram parameters and values for the lower and upper bounds for the curve fitting, the exponential variogram displayed in Figure 5 has been chosen, as it yields the most sensible results in the final kriging map.

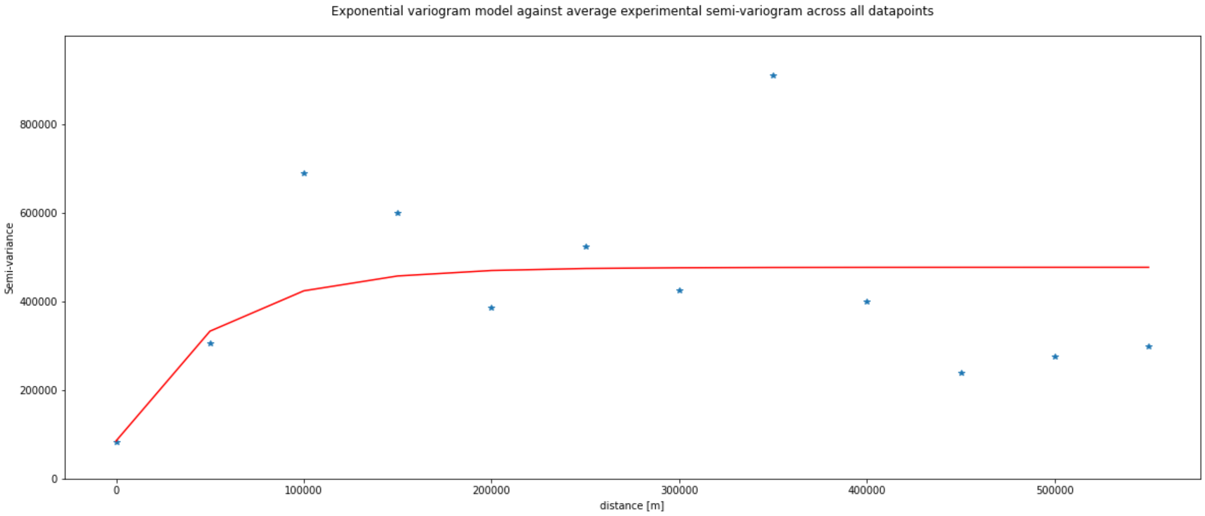

Figure 5: Average experimental semi-variogram and fitted exponential variogram model

Figure 5 shows the average of the experimental semi variograms from all origin-datapoints as the blue stars and the fitted exponential variogram model as the red line.

The variogram model has a range of 150,000, sill of about 390,000 and a nugget of about 85,000.

Even though with some different parameters the gaussian variogram resulted in a better fit for close ranges, it resulted in a lot of insensible results in terms of negative, or extremely high positive values during the kriging process.

2.2.6 Kriging

A function for Ordinary Kriging is defined, returning the kriging estimate and kriging error.

A kriging grid is defined. Through a trial-and-error process a grid size of 100 x 100 grid cells has been chosen as the final grid size to achieve a sufficient resolution, while a smaller grid size of 20 x 20 grid cells has been used to speed up trial-and-error process on the experimental variogram and variogram modelling parameters, sacrificing resolution for a faster turnaround time. The 100 x 100 grid takes about nine minutes to calculate while the 20 x 20 grid takes about 20 seconds.

The distance from each grid point, acting as the origin, to each datapoint is calculated. Then the Ordinary Kriging function is applied to each grid point using the distances between the datapoints, the distances from the datapoints to the grid point and the original values for the expected production at the datapoints.

Insensible values below 0 are adjusted to 0 and the results are slightly rescaled to fit the existing datapoints.

This results in a kriging estimate map and a kriging error map that are visualized with the existing datapoints over the windspeed and terrain maps.

3 Discussion and Results

3.1 Spatial Data Analysis

The most important results from the spatial data analysis are presented and discussed in the following part. Further plots, scatterplots and histograms can be found in the accompanying jupyter notebook.

3.1.1 Wind & Terrain Complexities Map

Figure 6 shows the location of windfarms in Norway, color-coded to the expected energy production in GWh, over a map of the average windspeed at 50m above ground in m/s.

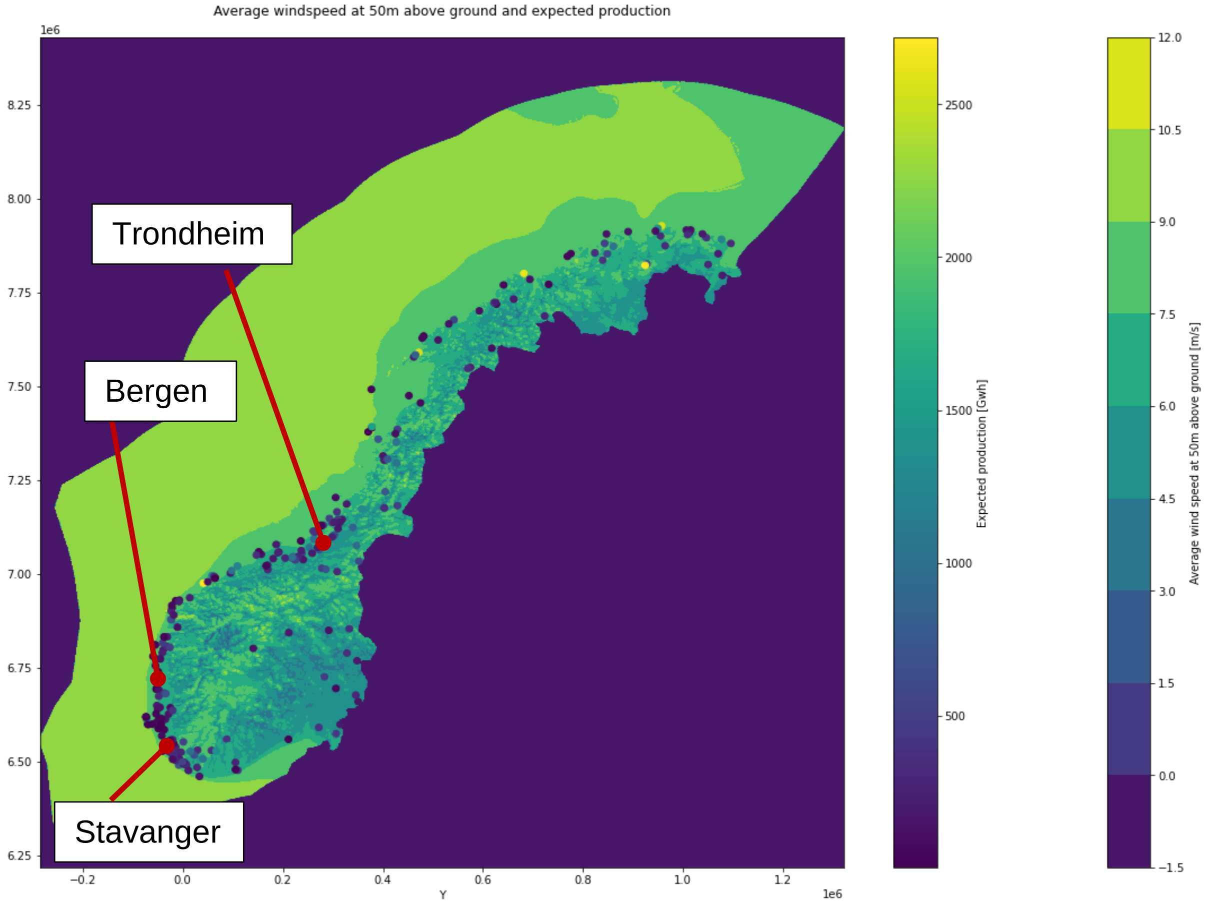

Figure 6: Map of Windfarms in Norway over average Windspeed

From Figure 6 it is clearly visible that most of the windfarms are located along the Atlantic shoreline, where there are consistently high windspeeds, with a higher windfarm density towards the south. Additionally, a big portion of Norway's population is situated along that shoreline including the 2nd, 3rd and 4th largest cities by population, Bergen, Stavanger and Trondheim, resulting in higher energy demand and better power-grid infrastructure than in the areas further inland.

Figure 7 shows the location of windfarms in Norway, color-coded to the expected energy production in GWh, over a map of the terrain complexity.

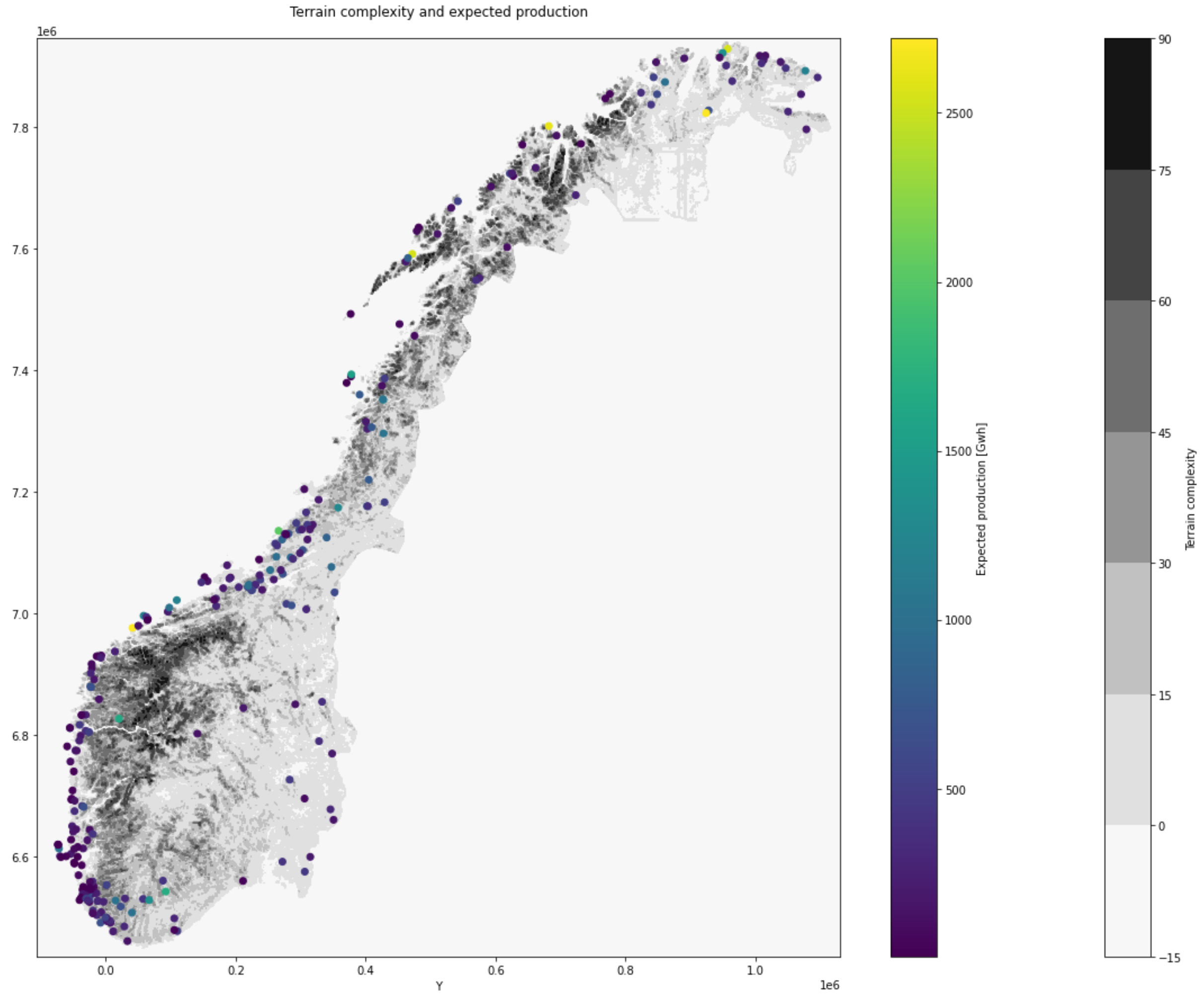

Figure 7: Terrain complexity and expected production

From Figure 1 and Figure 7 it is visible, that it is generally avoided to place windfarms directly in very complex terrain or leeward of it, as this would create turbulences in the Windstream, reducing the efficiency of the windfarm.

3.1.2 Histogram

Figure 8 and Figure 9 show the histograms on terrain complexity and windspeeds at the windfarm locations respectively.

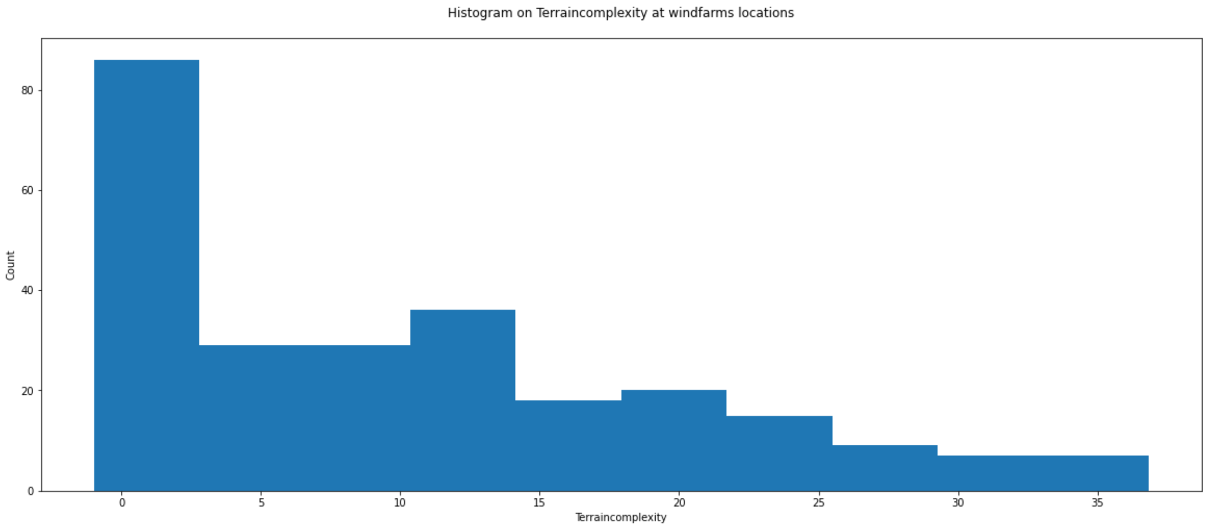

Figure 8: Histogram on Terrain Complexity at Windfarm Locations

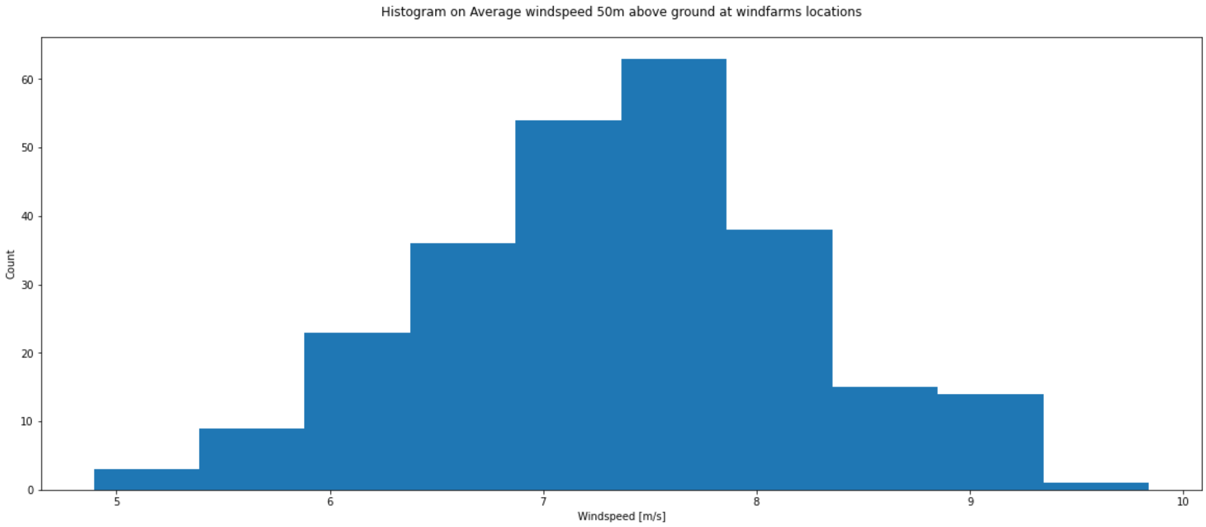

Figure 9: Histogram on Average Windspeed 50m Above Ground at Windfarm Locations

From Figure 8 we can see that most of the existing windfarms are located at low terrain complexities, while from Figure 9 it is visible that most of the windspeed data is in the range from 6.5 - 8.5 m/s following a normal distribution.

Figure 10 shows the histograms on licensed power in MW and expected energy production in GWh of the windfarms.

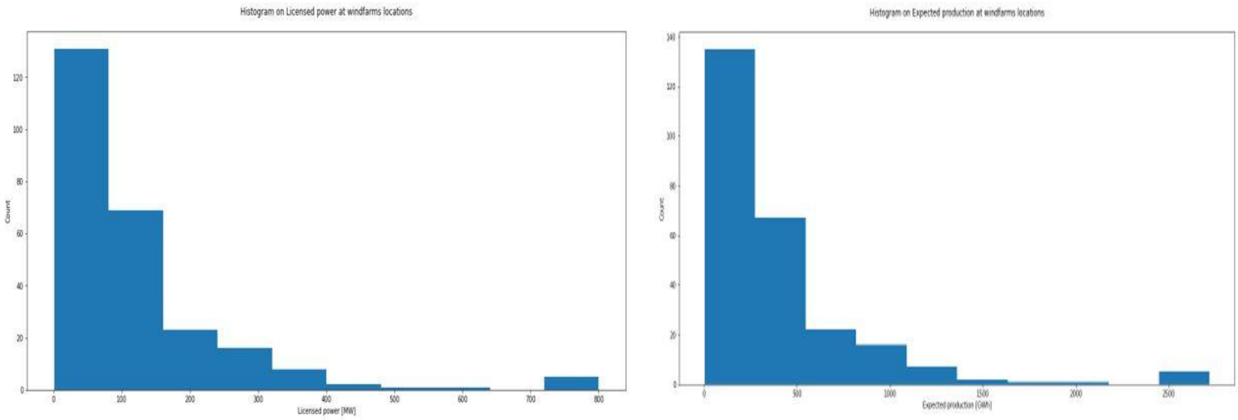

Figure 10: Histogram for Licensed Power and Expected Production

From Figure 10 we can see that most of the windfarms have a licensed power below 300 MW and expected energy production below 1000 GWh. It is also visible, that licensed power and expected production follow the same distribution.

3.1.3 Scatterplots

Figure 11 shows a scatterplot on the licensed power and expected production of the windfarms.

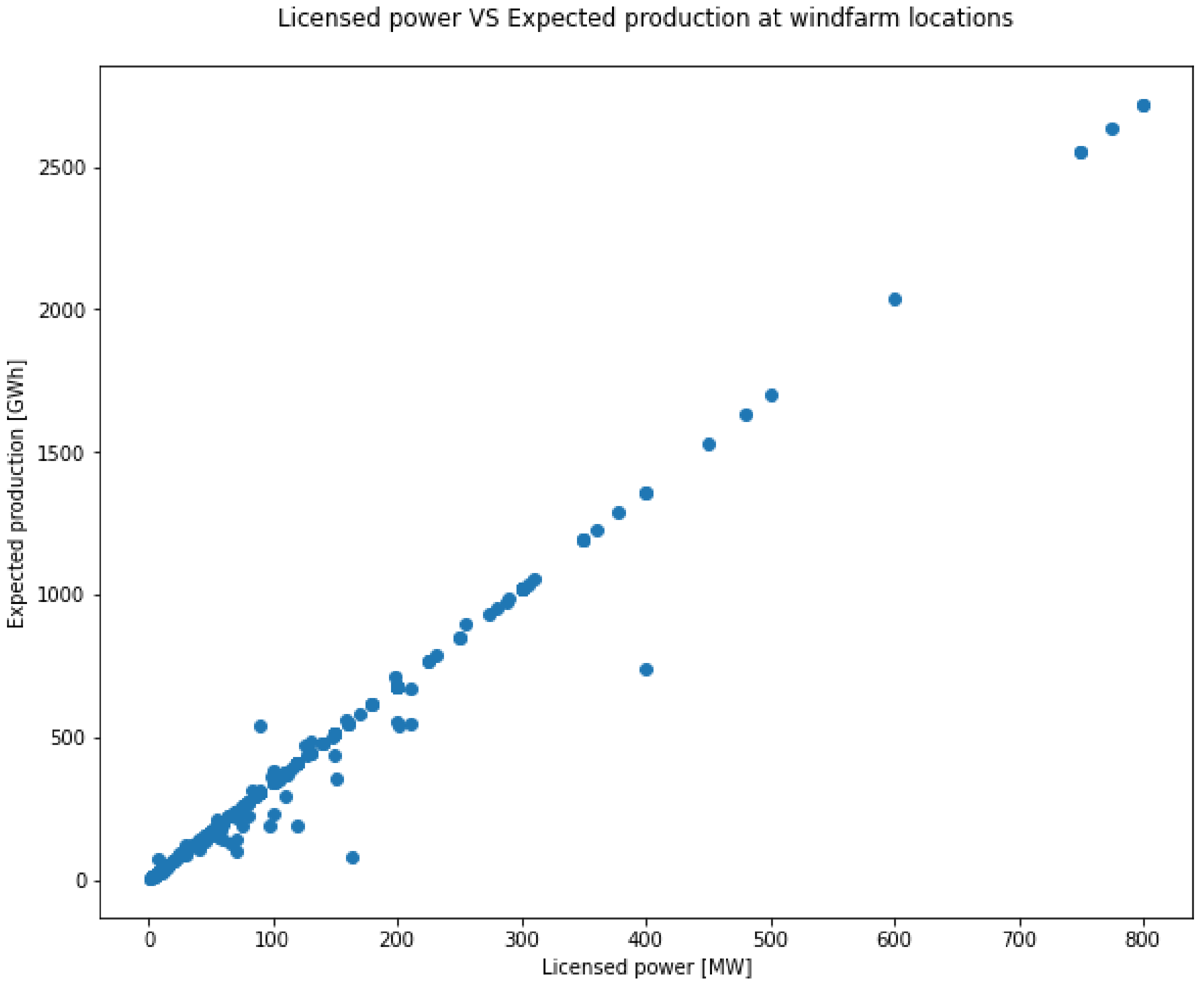

Figure 11: Scatterplot on licensed power VS expected production

From Figure 11 it is clearly visible that there is a linear correlation between the licensed power and expected production of a windfarm, with only a few outliers. This makes sense, as even under the best circumstances a low-power wind turbine can not produce more energy than the technical limitations of the turbine allow. While it is possible for a high-power turbine to produce low amounts of energy under suboptimal circumstances this plot shows, that the licensed power, which is given by the effective power of a single wind turbine through the technical design of the turbine and the number of turbines in a windfarm, is generally matched pretty well to the conditions in which the windfarm is placed to get optimal energy production.

3.1.5 Construction Date

From Figure 12 it can clearly be seen that most of the windfarms have been constructed in recent years, which emphasize the focus on the transition towards wind energy in Norway.

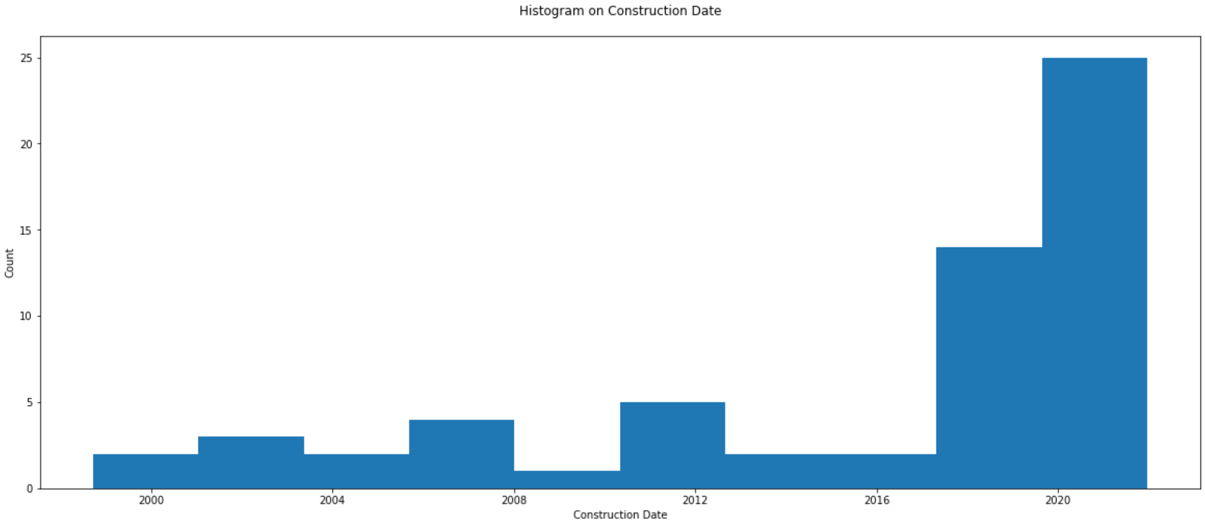

Figure 12: Histogram on Construction Date

Figure 13 shows the licensed power and expected production over the construction date of the windfarms.

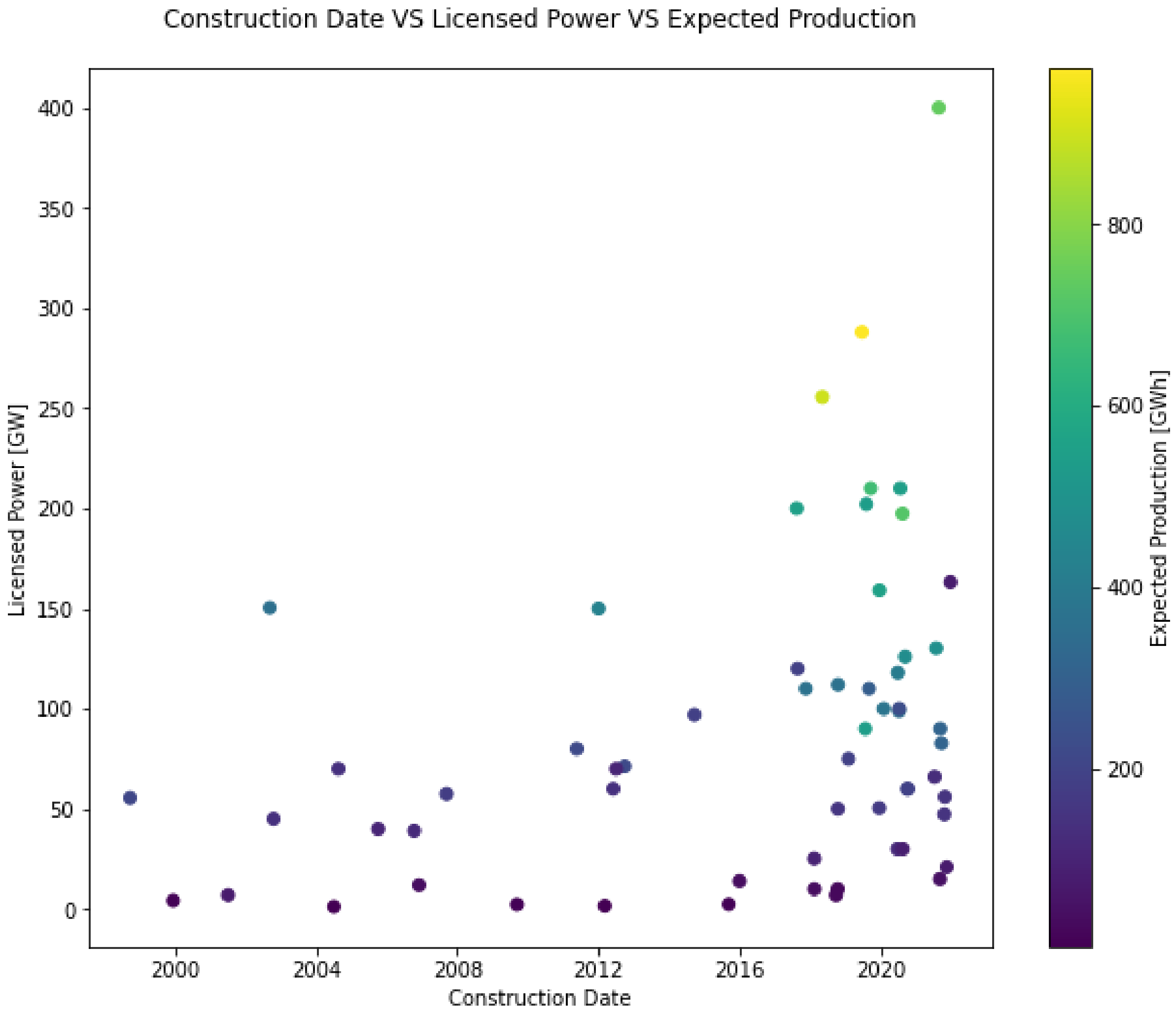

Figure 13: Construction Date VS Licensed Power VS Expected Production

From Figure 13 it is visible that high power windfarms with high expected production have only been constructed in recent years. It also shows that for the same licensed power the expected production is increased on windfarms from recent years, compared to earlier windfarms, indicating that the efficiency of modern windfarms has been increased.

3.2 Spatial Modelling

The most important results from the spatial modelling are presented and discussed in the following part. Further results can be found in the accompanying jupyter notebook

3.2.1 De-Clustering

Figure 14 and Figure 15 show the data map for the naïve data and the weighted data after the de-clustering respectively.

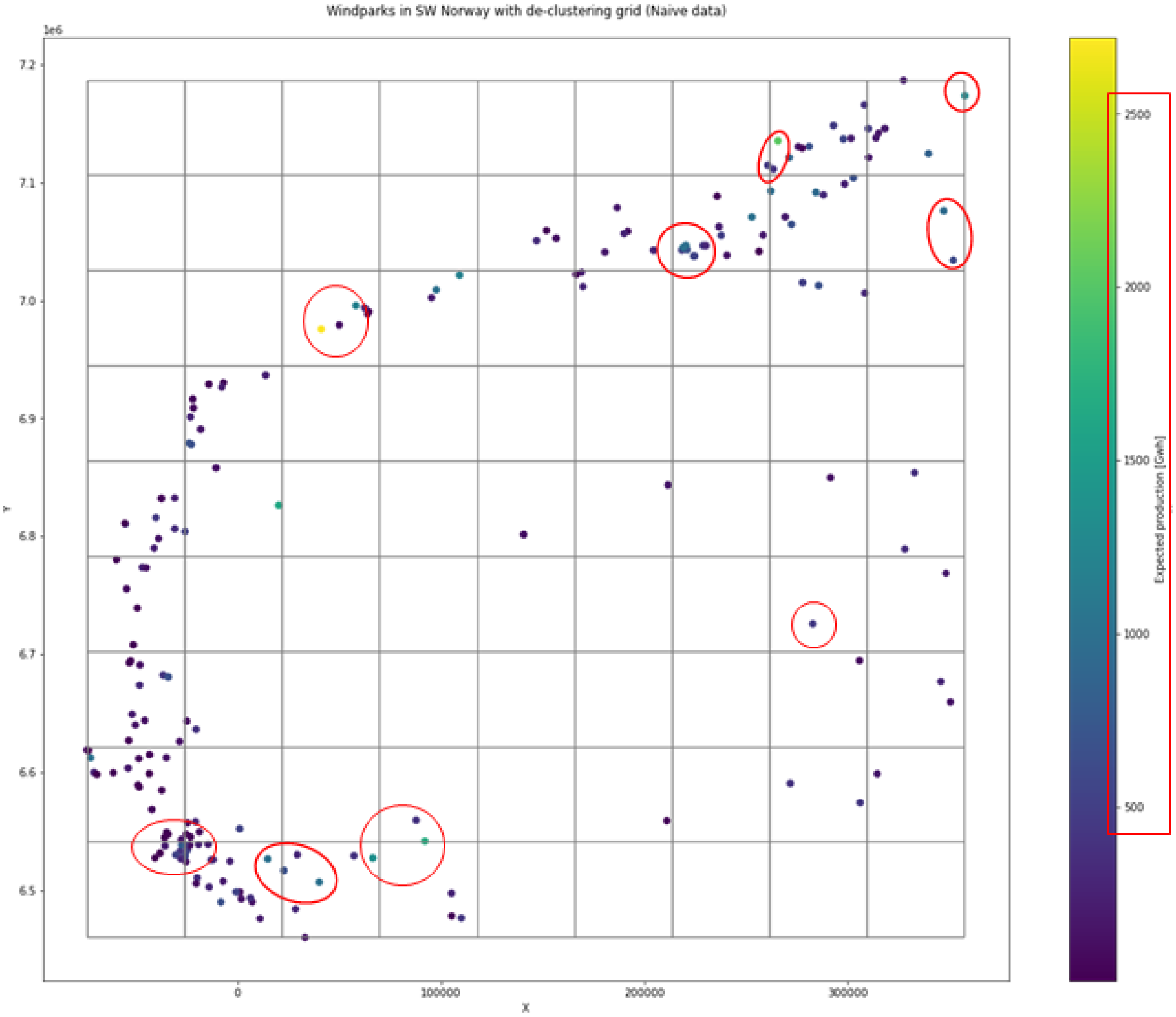

Figure 14: Naive Data Map

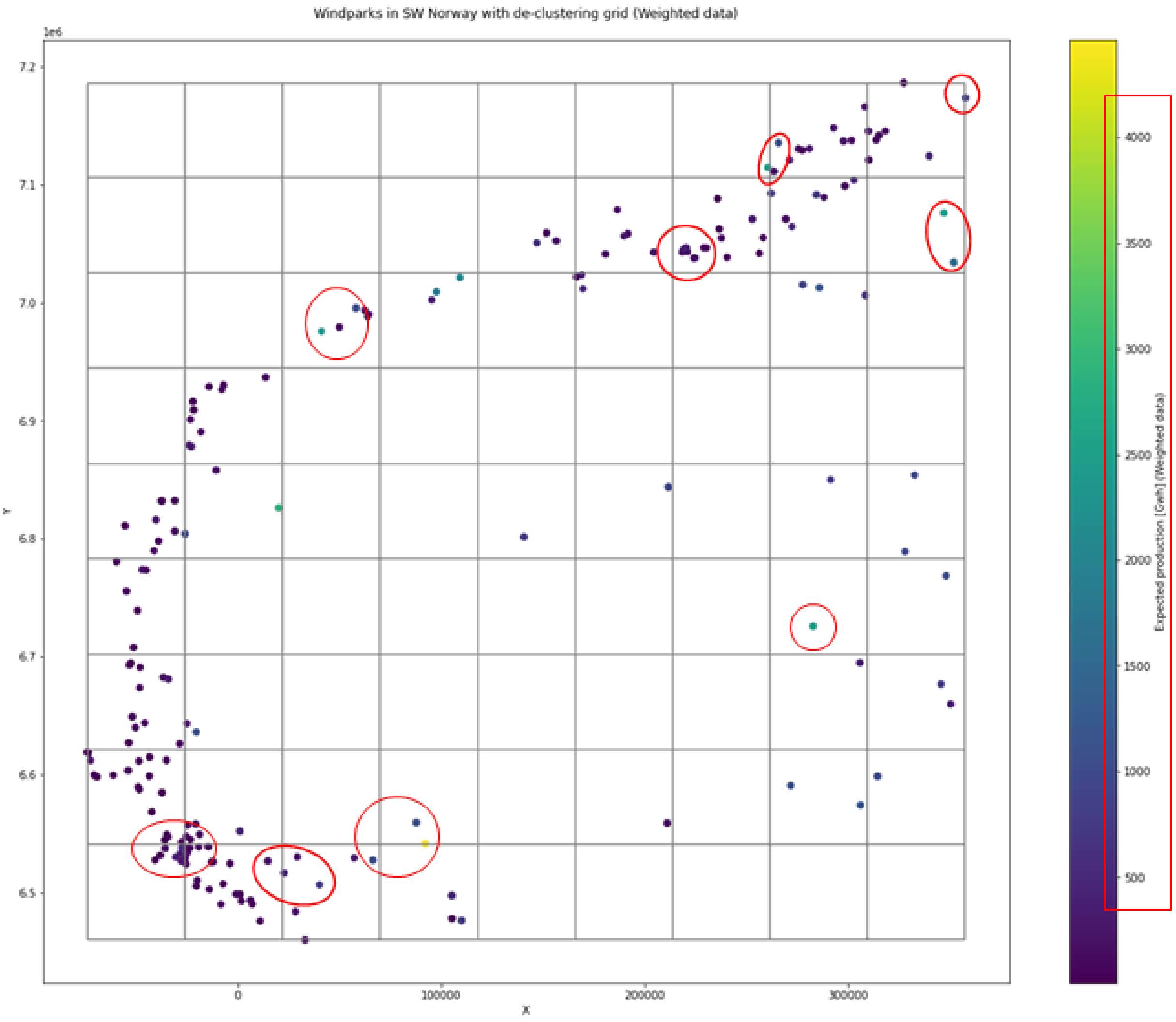

Figure 15: Weighted Data Map

Note that the scale for the expected production value has been increased in the weighted data, as some of the high production windfarms are in grid cells with low number of other surrounding windfarms. The change in relative energy production difference between neighbouring windfarms is visible through different colour changes among these neighbouring windfarms, which is visible in some of the clusters marked in red.

3.2.2 Kriging

The final map of the kriging estimates on expected energy production is displayed together with the expected energy production of existing windfarms over the terrain complexity map in Figure 16.

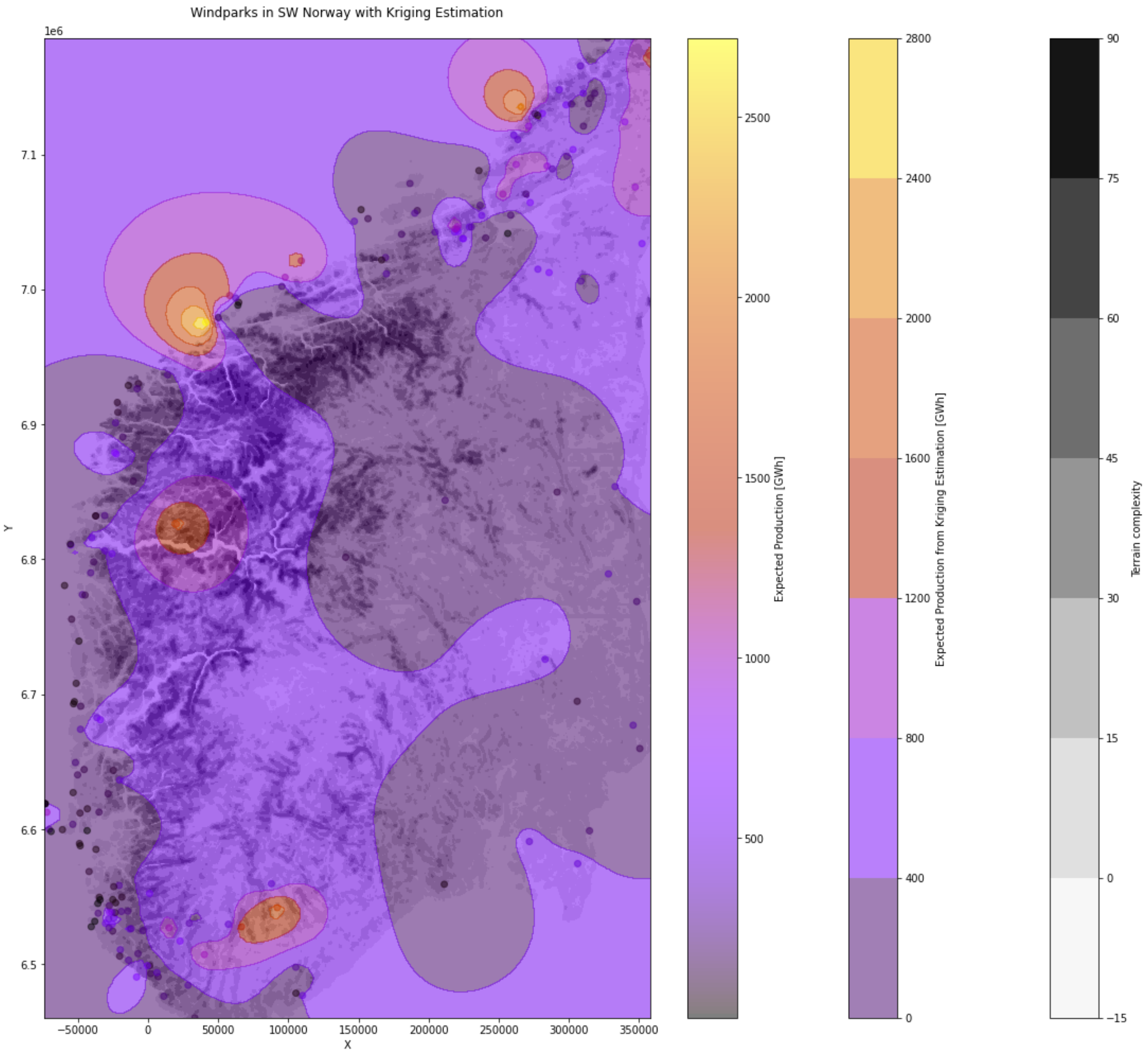

Figure 16: Final Kriging map

From Figure 16 it is visible that the kriging results in a good, smoothed interpolation between the values of the existing datapoints.

Figure 17 shows a comparison of our kriging map with the NVE national framework for wind power map.

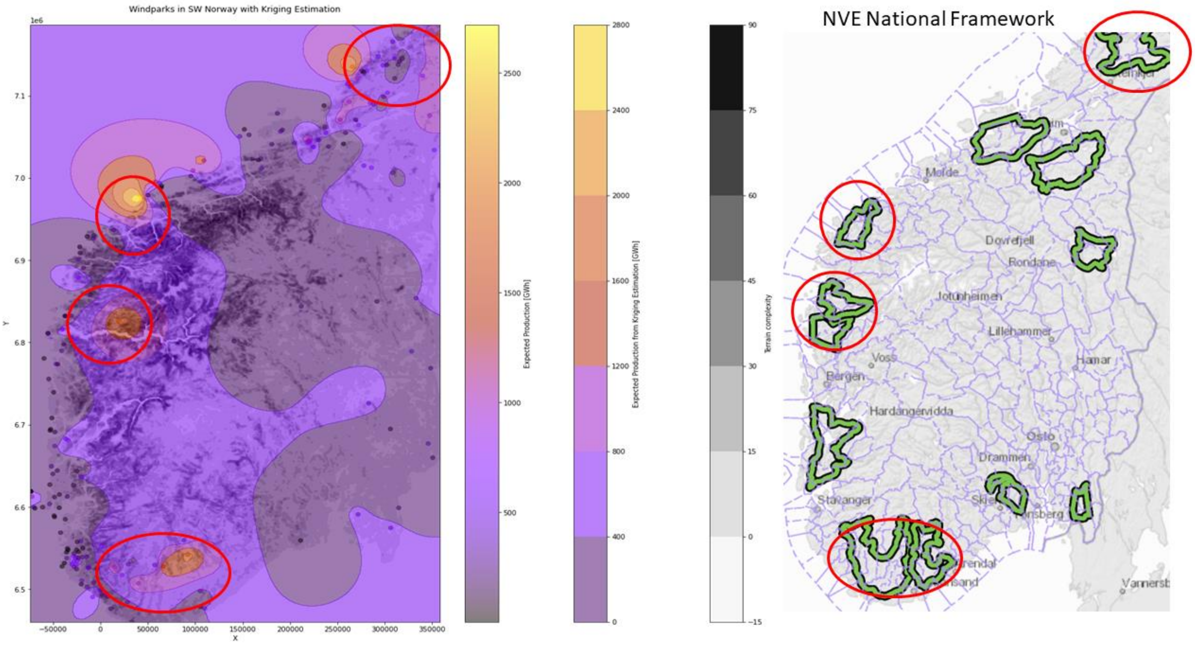

Figure 17: Comparison with NVE National Framework [2]

Figure 17 shows that some of the areas proposed by the NVE as the most suitable areas for onshore wind power, marked in red, are also roughly recommended by our kriging map.

Note that the kriging map does not suggest areas along the shoreline that intuitively would suit themselves for high energy production because of higher windspeeds and low terrain complexity, because there have not been a lot of high energy producing windfarms constructed there, but rather a lot of lower energy producing windfarms. The reasons for that can not be inferred from this data. While some of the windfarms in that area are of an older construction date, there are also a significant amount of newer windfarms in that area. Therefore the reasoning for why they have lower expected energy production is unclear from the data and can only be speculated upon.

Figure 18 shows a map of the kriging error together with existing datapoints over the terrain map.

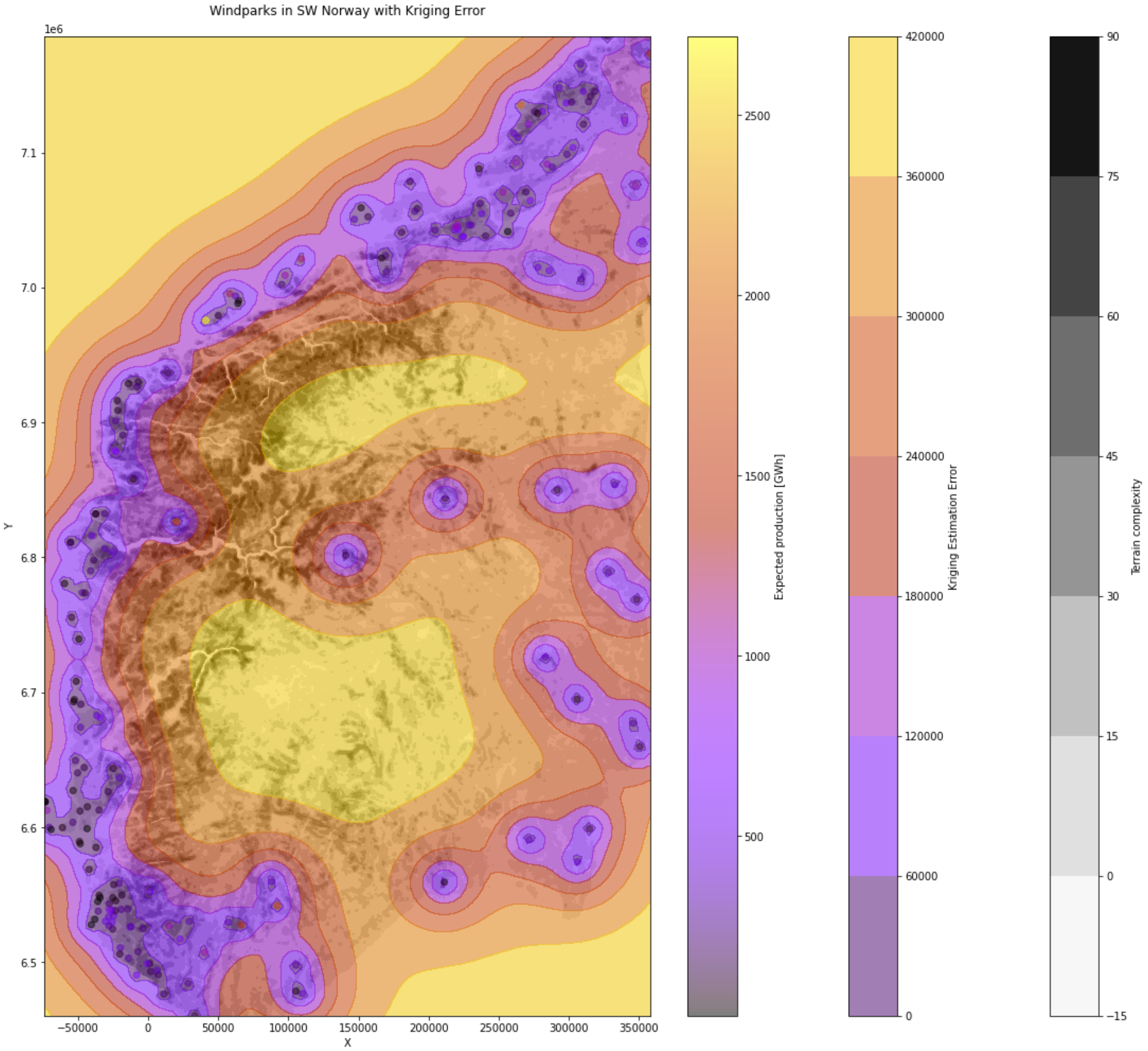

Figure 18: Kriging Error Map

From Figure 18 we can see that, as expected, the model has high confidence in areas where there is a lot of data leading to lower error values and lower confidence in areas with few or no datapoints leading to higher error values.

4 Recommendations and Conclusions

While intuitively the average windspeed plays a major role in finding a good location for a windfarm we saw from the spatial data analysis that it has not been the primary or only factor on determining a windfarms location and expected production, as windfarms can be found in medium and lower windspeeds as well and are not primarily placed in high windspeeds. This is because there are other factors to consider, such as the terrain complexity, which has a clearer correlation as in that high energy windfarms are primarily being placed in terrain with low complexity.

Other factors that might influence the decision on where to place a windfarm and designing the size and power of that windfarm, that were not discussed in this project, may include the proximity to residential areas, wildlife and airports for noise pollution, environmental and safety reasons, the local energy demand and the available power grid infrastructure and capacity.

Surely at least some of those factors have been taken into consideration in the past and are reflected in the data of the energy production and location of existing windfarms.

Applying the spatial modelling techniques discussed in this project results in a map that is purely based on that data and gives insight into which locations have been chosen for windfarms in the past and their expected production and, based on that, which areas might provide the potential for high energy production for future windfarms, but not necessarily why the locations have been chosen or why there is higher expected production in a given location. Also note that a large area with lots of lower energy producing windfarms, such as the south western shoreline, might in total produce more energy than a small area with a few high energy producing windfarms.

Another factor to consider is, as we saw in the section “3.1.5 Construction Date”, that the technology advances over time and allows for higher energy production under otherwise same circumstances. So, just because there are a lot of windfarms with low expected production in an area that doesn´t necessarily mean a more modern windfarm could not produce more energy in the same area.

To further improve the model more advanced kriging methods such as kriging with a trend or locally varying mean kriging with windspeed, terrain complexity and other data such as energy demand, grid capacity and population density as secondary data could be applied. With improvements to the computational efficiency or higher computational resources the model could be extended to cover whole Norway and, if the relevant data is provided, to other parts of the world.

Finally, it should be noted, that there are a lot of factors determining suitable and optimal locations for a windfarm that are not covered in this project as the authors had no prior domain knowledge in this area of expertise and approached this problem primarily from a limited, mostly data-driven, position, which can not cover all these factors.

References

[1] Meteoblue, "Simulated historical climate & weather data for Norway," [Online]. Available: https://www.meteoblue.com/en/weather/historyclimate/climatemodelled/norway_united-states_5004016. [Accessed 29 May 2022].

[2] The Norwegian Water Resources and Energy Directorate, "Map Services," Norwegian Government Agency, [Online]. Available: https://www.nve.no/map-services/. [Accessed 29 May 2022].

Appendix

A1: Jupyter Notebook