Assignment 1c

- Part 1
  - i
  - ii
  - iii
  - iv
  - v / vi / vii
    - A Note on Factor Correlation
    - A Note on Sample vs Population Data
    - Selecting the Sensitivity Analysis Tests to Perform
    - Product Moment Correlation Coefficient
    - Linear Regression
    - Variance Based Sensitivity Analysis
    - Notes on the sum of S_i and S_{Ti}
    - Sensitivity Analysis Limitation
- Part 2
  - Model 1 Analysis
  - Model 2 Analysis

Part 1

i

Complete, excluded from git for size

ii

Complete

iii

Five variables must be selected, can use judgement or research.

Will immediately focus more towards variables which might be useful for part 1b and future assignments - therefore variables such as `Grid Frequency`, and `Rotor Bearing Temperature` are disregarded.

Remaining suitable variables

Variable Name	Variable Description	Units	Notes
Ya	Nacelle_angle	deg	Irrelevant for own experiment
Ba	Pitch_angle	deg	Can be used in own experiment

Variable Name	Variable Description	Units	Notes
Wa	Absolute_wind_direction	deg	Could be influenced by control factors - e.g. deliberately facing across the wind when power is not needed
Ds	Generator_speed	rpm	Maybe possible to measure in own experiment
Rs	Rotor_speed	rpm	Appears to be ~1/10 the generator speed, suggesting a 10:1 gearbox before the rotor
Ws1	Wind_speed_1	m/s	
Wa_c	Absolute_wind_direction_corrected	deg	
P	Active_power	kW	A suitable power measurement, along with Apparent Power
Rm	Torque	Nm	Could be calculated in small scale, but with relative difficulty
Ws	Wind_speed	m/s	Although may not be possible to measure for small scale, could indicate a major factor
Nu	Grid_voltage	V	Irrelevant to turbine - around 700V
Pas	Pitch_angle_setpoint		Lots of NaN values, maybe not measured for this turbine, part of control system
Va1	Vane_position_1	deg	Limited data points
S	Apparent_power	kVA	A suitable power measurement, however Active Power applies more to DC as well as AC
Ot	Outdoor_temperature	deg_C	Likely has a small effect but difficult to control in small scale
Va	Vane_position	deg	Not included for turbine data
Cm	Converter_torque	Nm	
Ws2	Wind_speed_2	m/s	
Na_c	Nacelle_angle_corrected	deg	Irrelevant for own experiment
Va2	Vane_position_2	deg	Appears to be the same as Va1
Q	Reactive_power	kVAr	Less useful power measurement, especially when switching to DC small scale

Active Power, P, will be used as the output variable.

The remaining factors were chosen as:

1. Pitch Angle: Highly applicable to 1b, and should play a major role in output power.
2. Wind Speed: Although difficult to measure in small scale without an anemometer, it should play a large role in output power.
3. Generator Speed: Should provide an intermediate variable between external factors listed above, and the output power. Could help find possible influences such as generator efficiency variation.
4. Torque: A fairly common measurement when determining power of a system, could help find additional influences similar to Generator Speed.

5. Outdoor Temperature: Likely has a much smaller effect than the other factors, however if an effect is present it should be fairly repeatable. Could help determine real-world positioning of wind turbines, e.g. in cold vs hot climates with similar wind conditions.

The average values are used for all, as this should be less influenced by outliers than Max and Min values, and may be more representative of the time period than a single Max / Min value.

iv

First data cleaning is applicable for all variables:

1. Remove any data points which are not for wind turbine `R80790`. This should not remove any, but there is a chance that a complete different turbine's data is included, which would yield completely different results.

   ```
   % Remove any data points which are not the same wind turbine
   data = data(data.Wind_turbine_name == "R80790", :)
   ```

2. Remove any rows which contain missing data. Missing data points may have unexpected effects when conducting a sensitivity analysis, and could be indicative of further data corruption such as a failed network transmission. There are enough data points to limit concern with removing data.

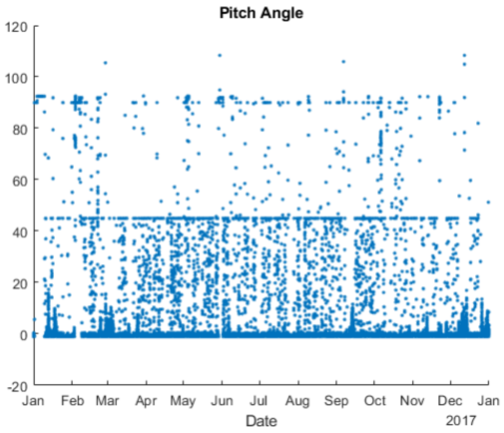
   ```
   % Remove any rows which contain missing values
   data = rmmissing(data)
   ```

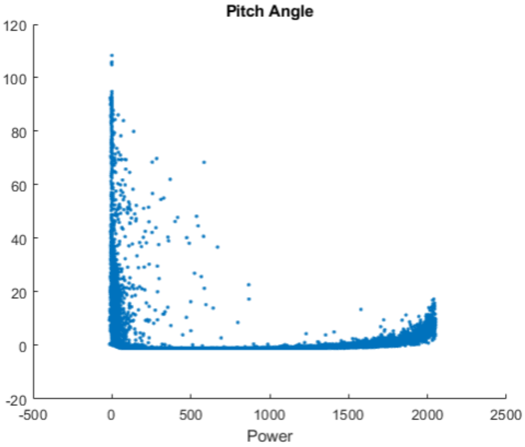
3. The `Date_time` string column is converted to a `datetime` object which can be used to plot in chronological order.

   ```
   % Convert date column to datetime objects
   data.Date_time = datetime(data.Date_time,"InputFormat", "yyyy-MM-dd HH:mm:SSz", "TimeZone", "UTC")
   ```

Then more manual checks were performed. A visual check may indicate a particular kind of outlier which should be removed. All variables were plotted on a time axis.

Pitch Angle





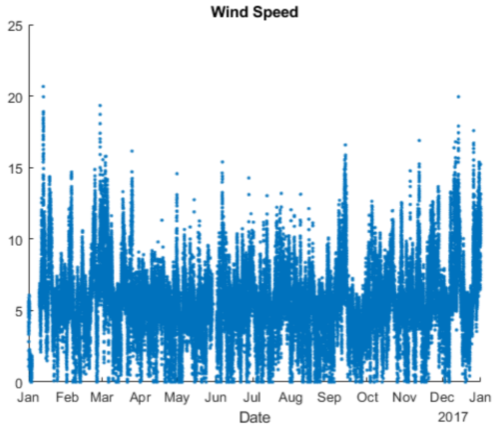
Some features immediately stand out - there is a visible line at 45°, suggesting this angle is used commonly, however there is no obvious correlation between 45° and high power, indicating this may be a default value when the turbine is not in use. The same can be said for 0°.

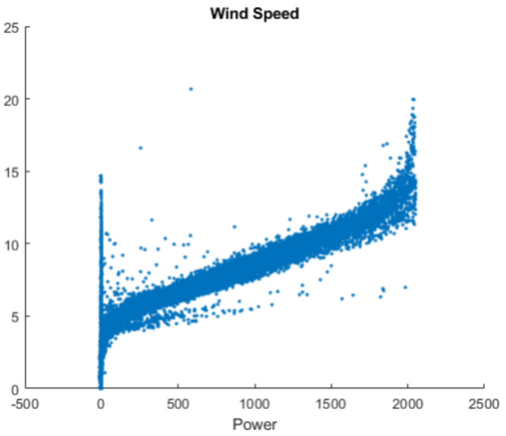
Few values appear above 90° - although specific information about the turbine could not be found through research, it can be assumed that the blades are not supposed to rotate past 90°, and therefore that these values are outliers. These values are removed.

```
% Remove data with pitch angles greater than 90
data = data(data.("Pitch Angle") < 90, :)
```

There is a large amount of data at low power and high angle. This could be assumed a result of operator control over the machine, for example increasing the pitch angle to prevent the turbine self-starting when power demands are met. However, as detailed information about the turbine or operating procedures could not be found, this assumption cannot be confirmed. Therefore, it is not possible to confidently remove this data as outliers, and therefore it will be left in the dataset.

Wind Speed



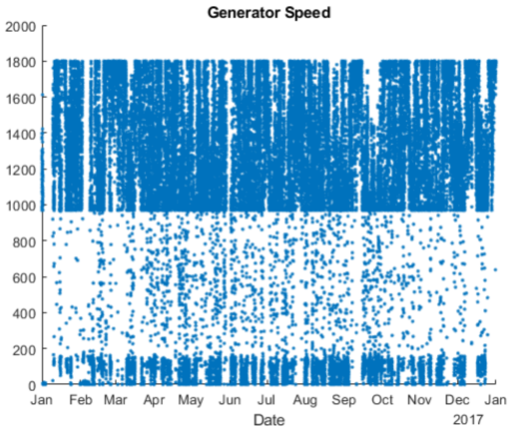


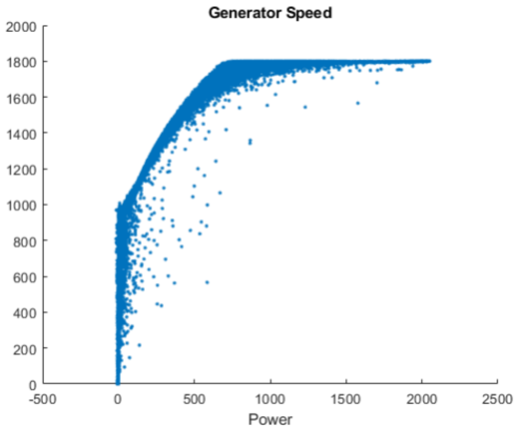
Removing data points that are a result of the control system, rather than the turbine physics, are recommended. It can be seen that there are a lot of data points where wind speed is $\neq 0$, however power is 0. This could be indicative of the turbine being stopped when electricity demand is already met. Therefore, all data points where $Wind\ Speed > 5\ \&\ Power < 50$ are removed.

```
% Remove data with with low power and not low wind speeds
data = data(~(data.("Wind Speed") > 5 & data.("P_avg") < 50), :)
```

Although a similar issue to removing data points from high pitch angles at low power, it can be assumed with enough certainty that the turbine should never have low / no power at moderately high wind speeds, and the only reason for this data must be operator control.

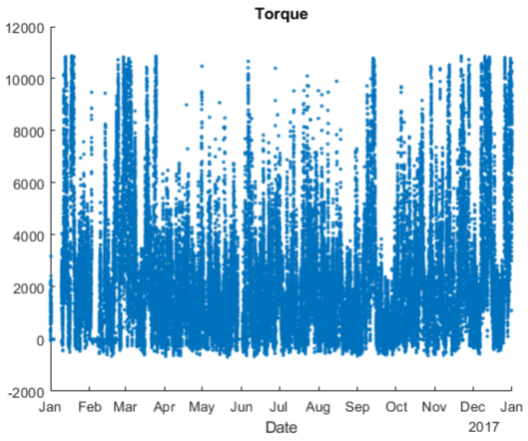
Generator Speed

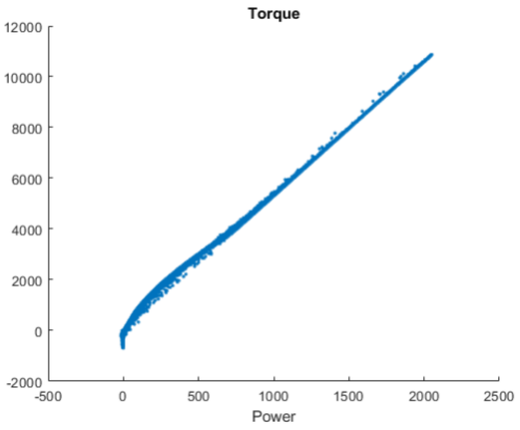




It could be argued that generator speeds below 1000 are as a result of system control factors - however there is not enough information known about the turbine to justify this, and therefore all the data will be kept.

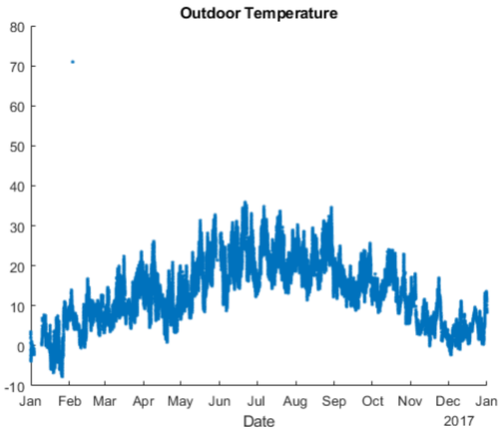
Torque

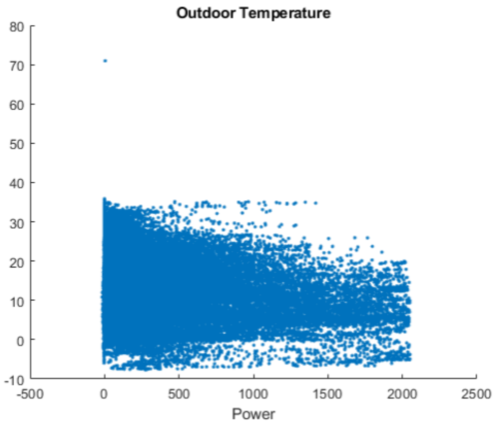




Appears to be a very obvious positive correlation, with no noticeable outliers. Therefore, the data will not be changed.

Outdoor Temperature





It is clear to see a single outlier, where temperature is ~71°. As this is higher than the highest officially recorded temperature on Earth of 56.7 °C, it can be assumed it is an outlier, and the data point is removed.

```
% Remove data with temperature higher than 50
data = data(data.("Outdoor Temperature") < 50, :)
```

Re-plotting the graphs and a visual check confirms that the mentioned outliers have all been removed.

v / vi / vii

A Note on Factor Correlation

It is understood that some of the factors may be correlated in some manner, for example, torque and generator speed. However, in part due to the lack of understanding of the physical turbine, the selected factors will be accepted for sensitivity analysis.

A Note on Sample vs Population Data

Because the data points are all time-based (e.g. more data can be obtained by simply querying at a new time), and time is always increasing, it can be said that the data is sample-based. It is not possible for the data to contain all possible data points, as is required for population data.

Selecting the Sensitivity Analysis Tests to Perform

1. OFAT is often the easiest when conducting experiments with cheap data acquisition, however in the case of the supplied data set, this is less useful. To 'hold all other factors constant' would require searching for data points where all factors (excluding the one in focus) are effectively the same, and with real-world data and four different factors, this is impractical.
2. Pearson Correlation Coefficient (I learned as Product Moment Correlation Coefficient, and will refer to PMCC) is an extremely easy step to perform on each of the data sets. In conjunction with the scatter plots from part iv, this will be performed on all the data. A major limitation is that this analysis only looks at interaction between one input and one output variable.
3. Linear Regression (LR) is harder to implement if multiple inputs are to be considered. It is likely more suited to data with a vaguely linear correlation. Using the scatter plots, Wind Speed and Torque likely have some form of linear correlation with Power, and Generator Speed and Pitch Angle may have linear sections, whereas Outdoor Temperature is more questionable. Linear Regression analysis will be performed on all factors excluding Outdoor Temperature, which will be left for more complex analysis.
4. Local Sensitivity Analysis using Partial Derivatives may be possible using tools such as DifferentialEquations.jl, however due to complexity and lack of existing information found while researching, this method will not be used.
5. Global Variance-Based Sensitivity Analysis will likely perform better with the data than PMCC due to interaction between variables, and LR due to ability to determine higher order interactions.

Product Moment Correlation Coefficient

Using the standard formulas, MATLAB code was written which calculates the PMCC for all variables, against power. Although standard MATLAB functions exist for this purpose, the provided equations were used and then validated against the built-in functions.

```
% Calculate standard deviation for x (Power)
N = length(data.P_avg);
x_mean = ones(N, 1) * mean(data.P_avg);
s_x = sqrt(sum((data.P_avg - x_mean).^2) / (N - 1));  % Could also be calculated with std function
r_xy = zeros(length(variables_named), 1);

% Iterate over all variables
for i = 1:length(variables_named)
    % Calculate standard deviation for y
    y_mean = ones(N, 1) * mean(data.(variables_named(i)));
    s_y = sqrt(sum((data.(variables_named(i)) - y_mean).^2) / (N - 1));

    % Calculate PMCC, could also use corrcoef
    t_x = (data.P_avg - x_mean) / s_x;
    t_y = (data.(variables_named(i)) - y_mean) / s_y;

    r_xy(i) =  sum(t_x .* t_y) / (N - 1);
end
```

PMCC Results

Variable	Standard Deviation	Sample Correlation Coefficient
Pitch Angle	20.1983	-0.3455
Wind Speed	2.555	0.9183
Generator Speed	562.1054	0.7207
Torque	2380.7	0.9955
Outdoor Temperature	7.9845	-0.2371

On visual inspection, it appears that Wind Speed and Torque are both positively correlated with Power, as expected.

If we assume a typical p-value of 0.05, we can test if our sample correlation coefficients are statistically significant enough to suggest that the population correlation coefficient is not zero; i.e. that there *is* a correlation. The p-values were calculated for all the results, however due to the very large sample size they were all very small, meaning the null hypothesis H0 (that there was no correlation between the variable and power output) can be rejected for each of them.

Wind Speed and Torque both show a strong positive correlation with the output power, which is to be expected as seen by visual inspection of the scatter graphs.

Linear Regression

The code for performing the linear regression and R2 is shown below. All the recorded data (input and output) is standardised to have a standard deviation of 1, and is centred around the mean value.

While the model is still a representation of the relationship between the input data values and the output data values, by standardising the data first, comparisons relating to variance between factors can be made. For example, comparing a b value of 1 to a b value of 0 shows that the first factor has a greater impact on the output variance of the output factor, than the second factor. A b value of zero could suggest that the input factor has no effect on the output factor variance. A b factor of one suggests the variance of the input factor will result in the same amount of variance in the output factor.

```
% Standardise model output
y_mean = ones(N, 1) * mean(data.P_avg);
s_y = sqrt(sum((data.P_avg - y_mean).^2) / (N - 1));
Y = (data.P_avg - y_mean) ./ s_y;
X = zeros(N, k);

% Iterate over all variables to create required matrix
```

```
for i = 1:k
    % Calculate standard deviation for y
    x_mean = ones(N, 1) * mean(data.(variables_named(i)));
    s_x = sqrt(sum((data.(variables_named(i)) - x_mean).^2) / (N - 1));

    % Standardise model inputs
    X(:, i) = (data.(variables_named(i)) - x_mean) ./ s_x;
end

% Calculate b values
b = (X' * X)^-1 * X' * Y
% Calculate R2
SS_tot = sum((Y - mean(Y)).^2)
SS_res = sum((Y - X * b).^2)
R2 = 1 - SS_res / SS_tot

% Calculate using built-in MATLAB function to check
[b, ~, ~, ~, stats] = regress(Y, X)
R2 = stats(1)
```

Linear Regression Results

Property	Value
b (Pitch Angle)	0.0228
b (Wind Speed)	-0.0425
b (Generator Speed)	-0.0648
b (Torque)	1.0978
b (Outdoor Temperature)	0.0131
R^2	0.9961

Due to the closeness of R^2 to 1, we can consider the data to be a good fit for the linear, additive model that is being fit to it.

From the results, we can see that variance of the input torque has the greatest effect on the output variance when compared to the other factors considered. In other words, varying the torque by e.g. 1 standard deviation, will result in the output power varying by approximately 1.1 standard deviations.

Meanwhile, the other factors have little influence over the output power variance according to the values obtained. This could be indicative of the output factor being less *sensitive* to these input factors; that they need to be varied a greater amount (in terms of standard deviation) than the torque to get a similar output power variance.

In reality, it may be the case that torque and output power are directly coupled, and that torque is less of an influence over power, and more of a direct conversion from one unit to another. This would explain why a small variance in the torque has a large effect over the output power, as they may be directly related. This large effect may be enough to cause the scale of remaining factors to be very small in comparison.

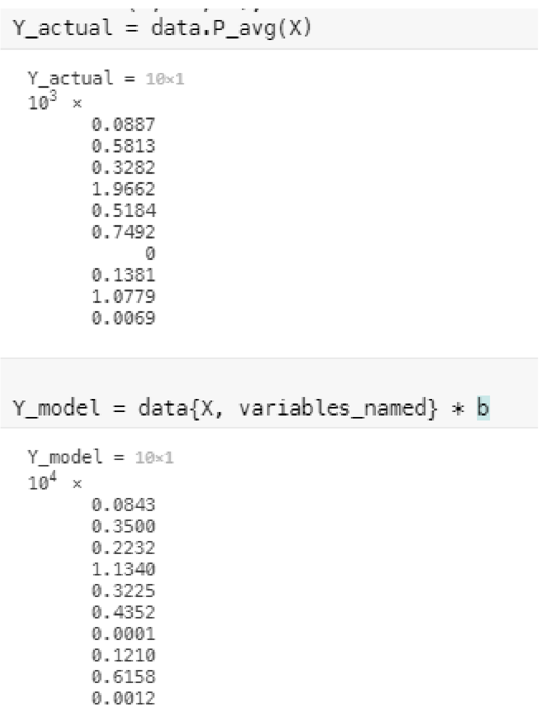
This could suggest that the effect of variance of the remaining factors on the output power variance is not necessarily very small, but that the effect of variance of torque on the output variance is much more significant in comparison.

To perform a quick check to ensure the results are sensible, a random selection of (standardised) input data points are fed back through the model, to determine if they have a similar result to the (standardised) actual recorded value. Due to the similarity between model and actual data points, this secondary check confirms the R2 value is likely accurate, and that no errors during calculation were made.

```
% Checking the model with 10 random points, to ensure model is in the correct ball park

% Y = bX
X = randi(N, 10, 1);
Y_actual = data.P_avg(X)

Y_model = data{X, variables_named} * b
```



```
Y_actual = data.P_avg(X)

  Y_actual = 10×1
  10^3  ×
       0.0887
       0.5813
       0.3282
       1.9662
       0.5184
       0.7492
            0
       0.1381
       1.0779
       0.0069

Y_model = data{X, variables_named} * b

  Y_model = 10×1
  10^4  ×
       0.0843
       0.3500
       0.2232
       1.1340
       0.3225
       0.4352
       0.0001
       0.1210
       0.6158
       0.0012
```

Variance Based Sensitivity Analysis

The provided equations for calculating the first order effect sensitivity, and total effect index were implemented in MATLAB.

```
% Calculate linear regression b with all data mapped between [0 1]
b = regress(rescale(data.P_avg), rescale(data{:, variables_named}))

% Create sobol quasirandom number matrices
sob = sobolset(2*k);
A = sob(2:N+1, 1:k);
B = sob(2:N+1, k+1:end);

A_B = zeros(N, k, k);
B_A = zeros(N, k, k);

% Create A_B, where A_B(:, :, i) contains the ith column of matrix B, and
% all other columns from A. Likewise for B_A.
for i = 1:k
    A_B(:, :, i) = [A(:, 1:i-1), B(:, i), A(:, i+1:end)];
    % B_A(:, :, i) = [B(:, 1:i-1), A(:, i), B(:, i+1:end)];
end
```

```
A_eval = zeros(N, 1);
B_eval = zeros(N, 1);
A_B_eval = zeros(N, k);

% Evaluate the model for all inputs in A, B, and A_B
parfor i = 1:N
    A_eval(i) = A(i, :) * b;
    B_eval(i) = B(i, :) * b;

    for j = 1:k
        A_B_eval(i, j) = A_B(i, :, j) * b;
    end
end

% Calculate the expected reduction and variance in the output, Y
V_X = zeros(5, 1);
E_X = zeros(5, 1);
for i = 1:k
    V_X(i) = mean(B_eval .* (A_B_eval(:, i) - A_eval));
    E_X(i) = 1 / 2 * mean((A_eval - A_B_eval(:, i)).^2);
end

% Not sure what input data is used to calculate Y maybe I can use the sobol
% random numbers - e.g. I can use A_eval?
Y = A_eval;
V = var(Y)

S = V_X / V
S_T = E_X / V
```

The resultant S and S_t values for each factor are shown in the table below.

Variance Based Sensitivity Analysis Results

Factor	S_i	S_Ti
Pitch Angle	0.1196	0.1196
Wind Speed	0.6976	0.6975
Generator Speed	0.0012	0.0012
Torque	0.0142	0.0142
Outdoor Temperature	0.1675	0.1674

The results are indicative of several things:

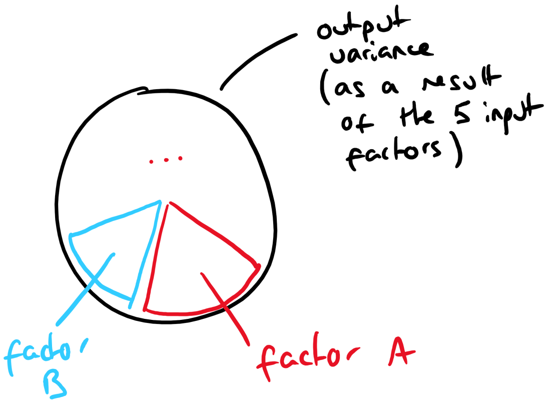
- Due to the small differences between S_i and S_{Ti}, this suggests the factors are fairly independent, and interactions between them do not contribute to the variance of the output.
- Wind speed is the most significant factor when compared to the other tested factors, for influencing the output variance.
- Generator speed is the least significant factor when compared to the other tested factors, for influencing the output variance. This could be because the generator speed may be directly linked to the output power, but does not influence it. For example, any variance difference between the speed and power could be caused by random noise, and therefore be very unrelated. It may be more suitable to think of generator speed as an output factor.
- A similar conclusion to generator speed can be drawn for torque.

- Outdoor temperature has a noticeable effect on the output power variance, based on the results of this sensitivity analysis.

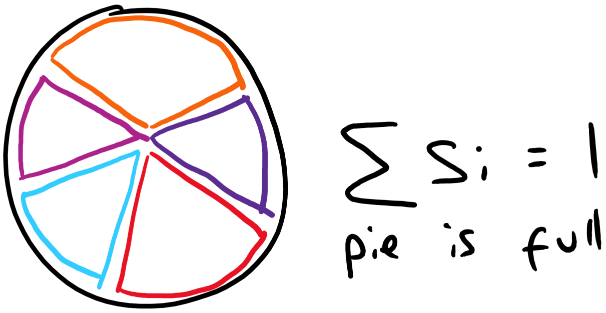
Notes on the sum of S_i and S_{Ti}

The variance of the output can be visualised as a pie chart, with the whole pie representing the total output variance explained by the five factors, and each slice representing the contribution to that output variance by each input factor.

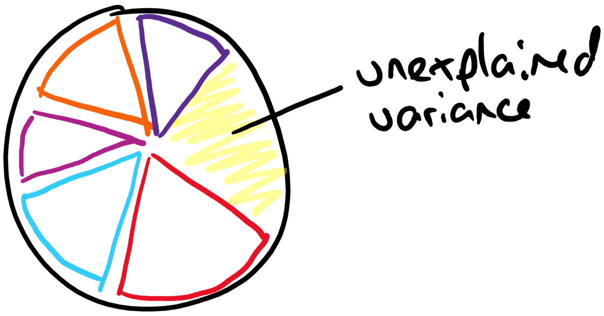
The S_i or S_{Ti} values represent the proportion of the pie that that factors slice occupies. In other words, a value of 0.5 means that input factor can explain half of the output factor variance.



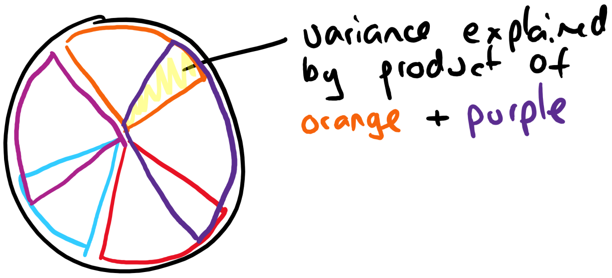
If $\sum S_i$ or $\sum S_{Ti} = 1$, this means that the variance of the input factors perfectly explain the variance of the output factor. They each contribute some proportion to the total output variance.



If $\sum S_i < 1$, this can be seen as a gap in the pie. There is some variance in the output which cannot be explained by variance in the input factors, when they are varied while all other factors are held constant. Therefore, it can be assumed that this unexplained variance must come from some interaction between factors, which was not allowed to happen for S_i as the other factors are held constant.



This 'leftover' variance can be seen in the pie chart of S_{Ti}, where overlap between the slices represents the variance explained by the interaction between two or more variables.

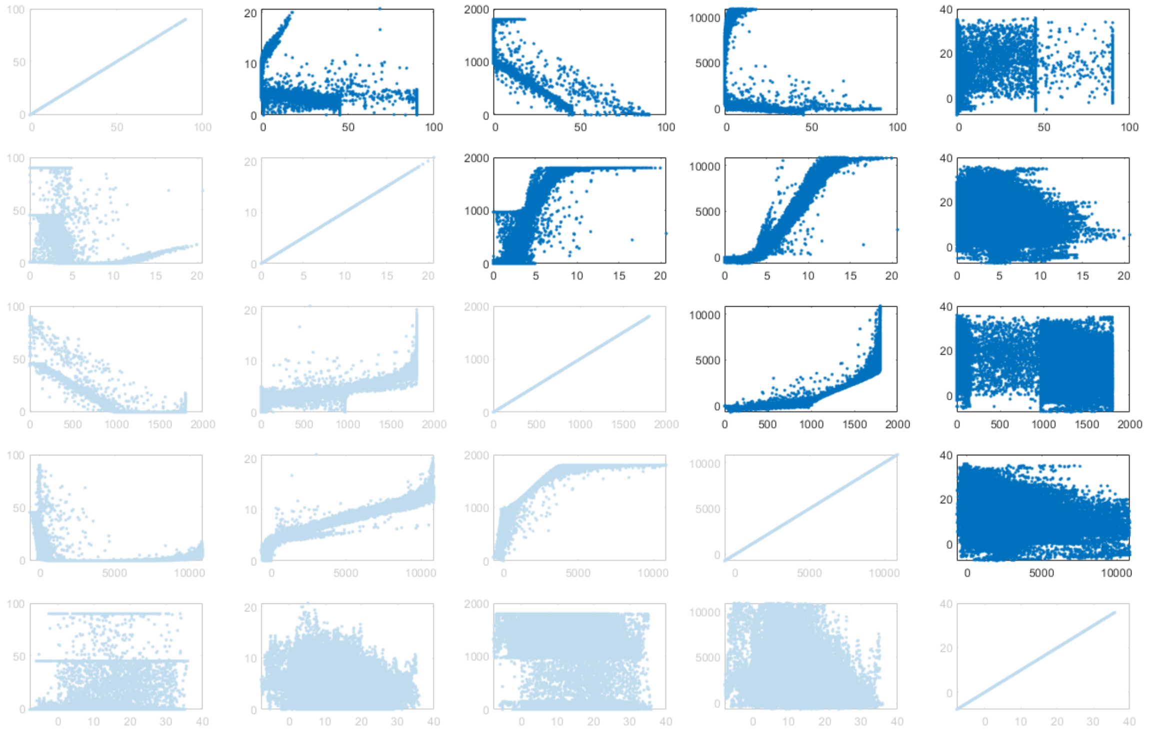


It can be seen that this overlap of explained variance should equal the missing section of the S_i pie. Mathematically, this means that $S_i + S_{Ti} = 2$, as together they create two whole pies.

Sensitivity Analysis Limitation

An important limitation of variance based sensitivity analysis is that variables must be independent from each other. If variables are strongly correlated, this can invalidate the method. To perform a quick check for variable correlation, scatter plots of the relationship between all possible factor interactions are shown below.

In the image, plots relating to the relationship between two of the same factor, and duplicate graphs, are ignored.

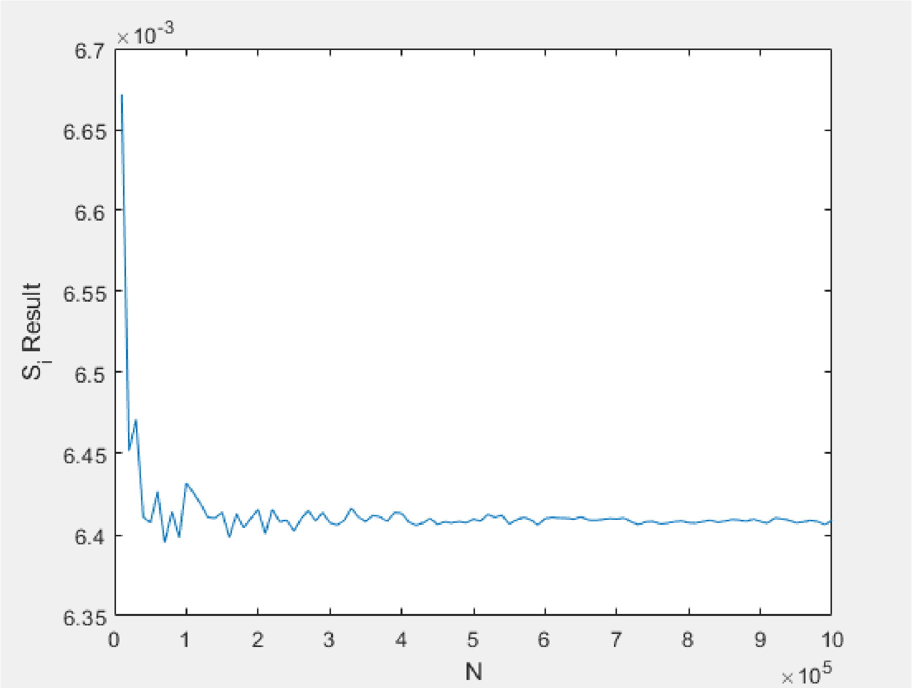


Visually inspecting the plots can help determine if there is a strong correlation between any two input factors. There is some correlation between some input factors, most notably at element (2, 4) (Wind Speed and Torque), which may promote caution regarding results obtained for these factors, however the correlation is not significant enough to warrant invalidation of the rest of the analysis.

Part 2

The code from part 1 was adapted to use the supplied confidential model.

To determine a sensible N, the code was run with N values in increments of 10000, from 10000 to 1000000. The output S_t values against N input values are shown below.



It appears that after ~400000 N, the results are approximately constant. Therefore, N = 500000 was used for testing.

The code is shown below:

```
k = 5
N = 500000
model = "1"

% Create sobol quasirandom number matrices
sob = sobolset(2*k);
A = sob(2:N+1, 1:k);
B = sob(2:N+1, k+1:end);

A_B = zeros(N, k, k);
B_A = zeros(N, k, k);

% Create A_B, where A_B(:, :, i) contains the ith column of matrix B, and
% all other columns from A. Likewise for B_A.
for i = 1:k
    A_B(:, :, i) = [A(:, 1:i-1), B(:, i), A(:, i+1:end)];
    % B_A(:, :, i) = [B(:, 1:i-1), A(:, i), B(:, i+1:end)];
end

A_eval = zeros(N, 1);
B_eval = zeros(N, 1);
A_B_eval = zeros(N, k);

% Evaluate the model for all inputs in A, B, and A_B
A_eval(:) = TurbineModel_2020(A, model, 13);
B_eval(:) = TurbineModel_2020(B, model, 13);

for j = 1:k
    A_B_eval(:, j) = TurbineModel_2020(A_B(:, :, j), model, 13);
end

% Calculate the expected reduction and variance in the output, Y
V_X = zeros(5, 1);
E_X = zeros(5, 1);
for i = 1:k
```

```
    V_X(i) = mean(B_eval .* (A_B_eval(:, i) - A_eval));
    E_X(i) = 1 / 2 * mean((A_eval - A_B_eval(:, i)).^2);
end

V_X
E_X

Y = A_eval;
V = var(Y)

S = V_X / V
S_T = E_X / V
```

The results are shown in the table below:

Confidential Model Results (Employee ID = 13)

Parameter	Model 1 S_i	Model 1 S_Ti	Model 2 S_i	Model 2 S_Ti
A	0.0064	0.0085	0	0
B	0.2924	0.6795	0.0123	0.0123
C	0.0018	0.0025	0.0003	0.0003
D	0.2932	0.6819	0.9377	0.9377
E	0.0155	0.0184	0.0497	0.0497

Model 1 Analysis

As the effect indices are different between first order and total effect, this indicates there are interactions between input parameters which will vary their effect on the output variance.

Parameter B and D contribute approximately equally to the output variance, and more than any of the other input parameters. When they are varied in isolation they contribute less than when they are varied in conjunction with other input parameters.

Parameter C contributes the least to the output variable, regardless of whether it is varied in isolation.

Model 2 Analysis

The first order effect indices, and the total effect indices are the same (to four decimal places). This indicates that varying each model parameter has the same effect on the output variance, regardless of whether the change is isolated, or in conjunction with other model parameters.

This could suggest there is no interaction between the model parameters, as varying one has no noticeable effect on the remaining parameters.

Relative to the other model parameters, parameter D has the highest contribution to the output variance when it is varied.

Conversely, parameter A appears to have no effect on the output variance.