Assignment 1a

> ⚠ All relevant code and files can be found at the below url.

XDGFX/design-optimisation

You can't perform that action at this time. You signed in with another tab or window. You signed out in another tab or window. Reload to refresh your session. Reload to refresh your session.

https://github.com/XDGFX/design-optimisation/tree/master/Assignment%201a

Initial Investigation

> Request 1: fit the greatest volume of liquid into the crate.
> Request 2: optimise the can shape to minimise material usage.

Assuming all cans are stacked on top of each other, and as many cans as possible are included, only two factors influence the volume stored:

1. Radius of each can
2. Height of each can

However, they are linked; as each can must store exactly 500ml. It is wasteful for the can to be larger than required, and so each can must have volume of 500 ml (stated that wall size is negligible).

The can material usage is determined by the surface area of the can; the surface area is covered in can material (e.g. aluminium), and it is assumed no material is used elsewhere.

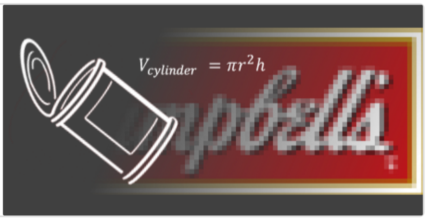
It is also assumed that material usage is constant regardless of where it is used - e.g. a square unit on the top or bottom of the can costs the same as a square unit on the side. This may not be true if there were different thicknesses of material, or processes used to create the shape.

Possibly Useful Notes

Optimal Can Dimensions

The purpose of a food can is to store food. It costs money to manufacture, store, and ship these containers. One would imagine, therefore, that over time a lot of thought has gone into their design and production.

http://datagenetics.com/blog/august12014/index.html



To minimise the surface area (material usage) to volume ratio, the height should be twice the radius.

The best known packings of equal circles in a square

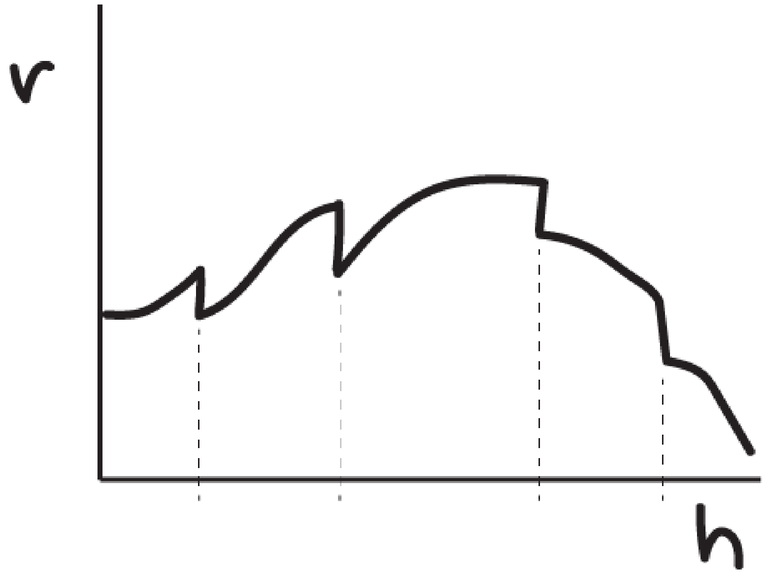
packing of equal circles in a square

http://hydra.nat.uni-magdeburg.de/packing/csq/csq.html

Optimal packing in 2D for circles in a square.

Computational Solution for Stacking Cans (Part iii)

There is some equation which models the volume of oil stored, with relation to radius and height. It may be complex as this is not a continuous equation, as once, e.g. height, reaches a specified value, the number of stacked cans will decrease by a discrete unit value.



The dotted lines indicate critical points in the function

This could likely be accounted for by using different equations for each *segment*, and having acceptable bounds.

Bounds

Assuming **all cans are stacked in an even grid**, the bounds can be found easily with the following equations which must be satisfied.

$$n * h \leq 560$$
$$n * 2r \leq 828.677$$

Where n is the number of cans in that given direction. Assuming cans can fit perfectly into the size given.

Because the box is square, and we are assuming an even grid, the total number of cans per level (in x and z) is interchangeable with the number of cans per side in the x and z dimensions. This will **not be the case if more complex packing arrangements are used**.

We could maximise the radius for each can n, and adjust the height to reach the 500ml volume per can. Alternatively, we could maximise height, and adjust radius.

If we vary radius within a fixed total number of cans on each level, there may be values at which a larger radius allows a smaller height, to a point that more cans can fit in the vertical direction. This will increase the total volume, if each can is still 500ml.

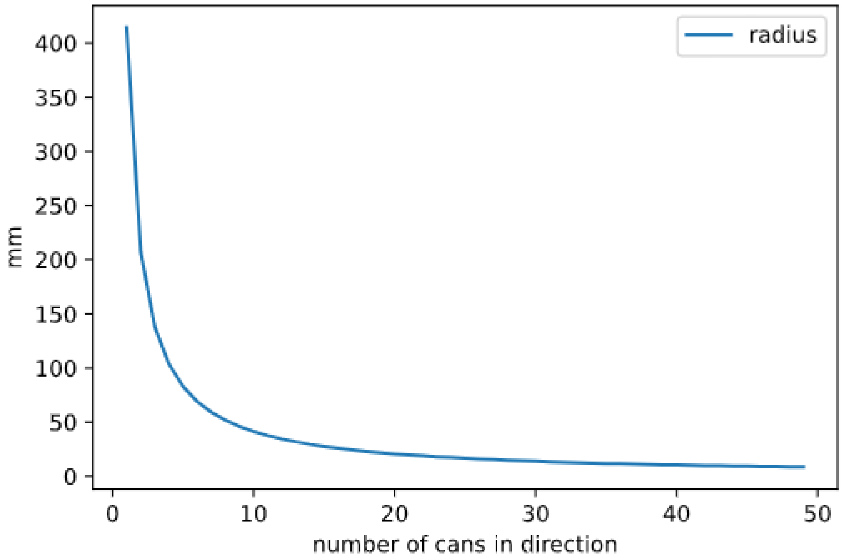
In other words, the greatest possible radius for a given number of cans per vertical level, will result in the smallest possible height per can, and therefore the greatest number of levels.

This can simplify the factors - as we only need to test the maximum possible radius for each n per level. Therefore:

$$r_n = \frac{828.677}{2n}$$

This can be converted into a 'lookup table' for different values of n. This was calculated up to 50 cans for each base dimension, and plotted for easier visualisation.

```
cans_x = range(1, 51)
crate_height = 560
crate_width = 828.677

max_radius = []

for n in cans_x:
    max_radius.append(crate_width / (2 * n))
```



We can then calculate the required height in order to reach 500ml volume

$$v = \pi * r^2 * h$$

$$h = \frac{v}{\pi * r^2}$$

$$h = \frac{500000mm^3}{\pi * r^2}$$

We can therefore calculate h for each value of r previously calculated. This can be used to calculate the maximum number of cans in the y direction, by using the previous formula:

$$n * h \leq 560$$

```
height = []
cans_y = []

for n in cans_x:
    required_height = 500000 / (math.pi * pow(max_radius[n-1], 2))
    height.append(required_height)
    cans_y.append(math.floor(560 / required_height))
```

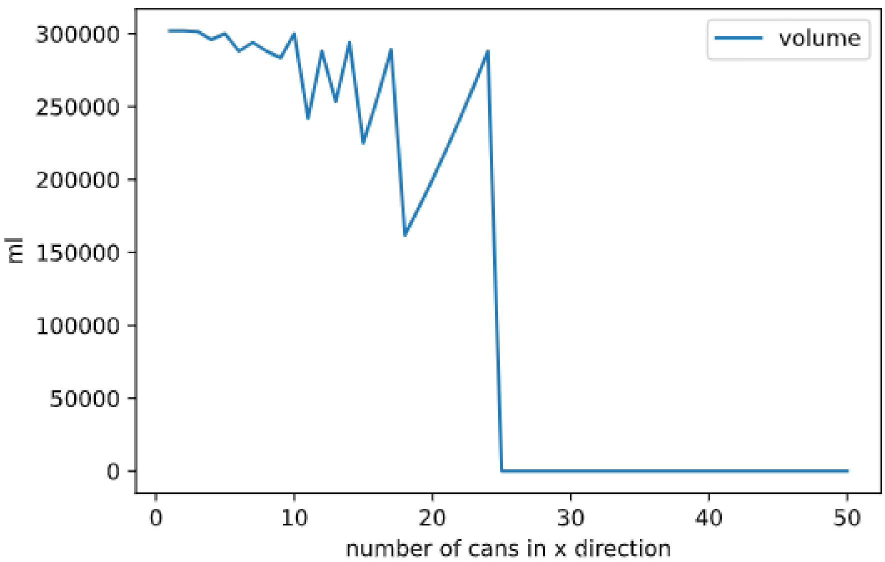
We can find the total cans with:

$$n = {n_x}^2 * n_y$$

And the resultant volume in ml with:

$$V = 500 * n$$

This produces the following graph:



Specifically, the scenarios where there is one can per x direction, and two cans per x direction (four per level) are the most efficient.

Theory 1

This is based on an even rectangular grid for cans.

1. Each can can be represented from the top-view as a square, with a percentage waste area (the corners outside the circle)
2. This waste ratio is always $1 - \pi * 0.5^2 = 0.2146$
3. Therefore, regardless of the number of circles in the square, the waste area is the same; the only waste change is area left on the sides of the square if the circles are not all touching (in this case they are).

4. This leaves waste **only** at the top, if the top level of cans does not touch the ceiling
5. With the largest radius possible, all circle sides are touching all square sides, so top down area is maximised. This also means the height is lowest, meaning this is likely the closest we can get to the top can touching the ceiling.

This explains why the volume gets worse with greater radius, and the oscillatory nature is explained by cans fitting well under the height, vs leaving a large air gap. The radius does not play a part in this case.

This can be converted to an analytical result by disregarding the top down view entirely;

- Stack n measurements of y within a total distance h.
- The value of y is based on the possible heights achieved by cans of radius $r = \frac{828.677}{2n}$, matching a volume $v = 500000mm^3$.

$$500000 = y * \pi * \frac{828.677}{2n}$$

$$y = \frac{1000000n}{\pi * 828.677}$$

$$y = 384.118n$$

We also know that:

$$y * n \leq 560$$

We can instead make both sides equal, as this will just introduce extra space in the top of each can, instead of the top of the crate. This will not affect the total number of cans.

$$y = \frac{560}{n}$$

Analytical Solution for Optimising Material Usage (Part i, ii)

To optimise material usage, surface area must be a minimum.

Therefore, the steps to find the optimal r and h values are as follows:

1. Determine an equation which links r and h based on a constant volume v. E.g. find h in terms of r.
2. Determine an equation which links r to the surface area a.
3. Differentiate this equation with respect to r to find the rate of change of surface area.
4. At $\frac{da}{dr} = 0$ this is a minimum or maximum surface area. These points can be found, and the minimum surface area can be found.

Based on previous working, we already know that:

$$h = \frac{500000mm^3}{\pi * r^2}$$

The surface area of a cylinder is $2\pi r(h + r)$ therefore:

$$a = 2\pi r(\frac{500000}{\pi r^2} + r)$$

$$a = 1000000r^{-1} + 2\pi r^2$$

And:

$$\frac{da}{dr} = -1000000r^{-2} + 4\pi r$$

At a minimum or maximum surface area, $\frac{da}{dr} = 0$.

$$0 = -1000000r^{-2} + 4\pi r$$

$$1000000r^{-2} = 4\pi r$$

$$\frac{1000000}{4\pi} = r^3$$

$$\sqrt[3]{\frac{1000000}{4\pi}} = r$$

$$r = 43.0127$$

Substituting back into the original equation gives:

$$h = \frac{500000}{\pi * 43.0137^2} = 86.0214$$

Combining The Solutions

The assignment states:

> They are now deciding how best to size these cans to minimise the amount of material they need for each can, and also to make their shipping more efficient.

This will be assumed to mean that can material optimisation is the primary objective, and shipping efficiency is secondary. Therefore, setting r and h to the values calculated, we get a maximum of 9.633 or $9\ cans$ in the x and z directions.

Using the previous calculations, this finds that a maximum of 7 cans are stacked vertically, resulting in 567 cans total at a volume of $286500mm^3$.

```
item = 8

print(cans_x[item])
print(cans_y[item])
print(max_radius[item])
print(height[item])
print(cans_total[item])
print(volume[item])

9
7
46.03761111111111
75.09215563384576
567
283500
```

Further Optimisation

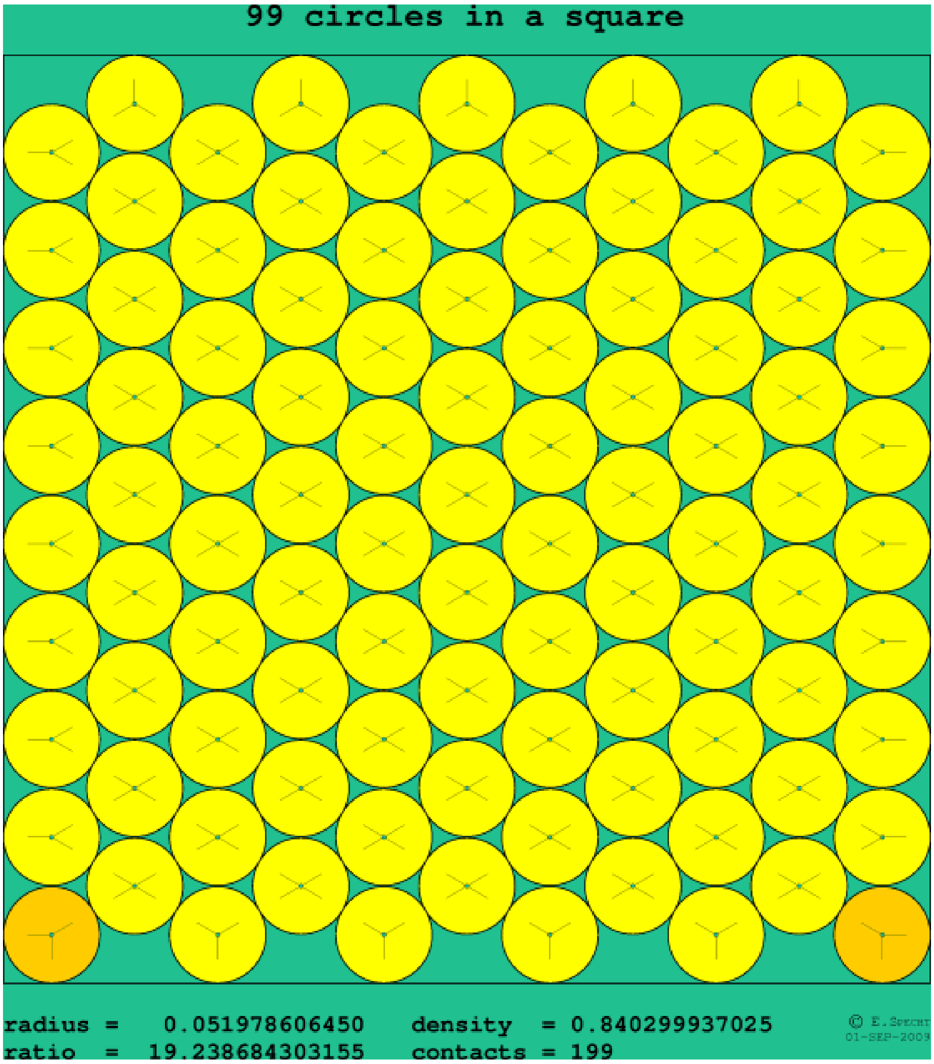
However, the grid pattern is not the most efficient arrangement of cans for each layer.

A table provided at http://hydra.nat.uni-magdeburg.de/packing/csq/csq.html compiles a list of optimal packing arrangements for equal-sized circles in a square. The square is of unit width and height, and the radii of each circle is provided.

Therefore, we can download this data and scale it proportionally to get a side length of 828.677.

```
import pandas as pd
import os

path = os.path.join(os.path.abspath(''), "circle_radius.csv")
circle_radius = pd.read_csv(path)

crate_width = 828.677
can_radius = 43.0127

# Scale all values to side length of crate
circle_radius["radius"] = circle_radius["radius"] * crate_width

# Then find the closest radius larger than or equal to our requirement
suitable_arrangements = circle_radius[circle_radius["radius"] > can_radius]

best_arrangement = [suitable_arrangements["cans"].iloc[-1], suitable_arrangements["radius"].iloc[-1]]
```

This gives the best arrangement of cans as 99 per layer (with a maximum possible radius of 43.0734, which means our smaller cans **will** fit).

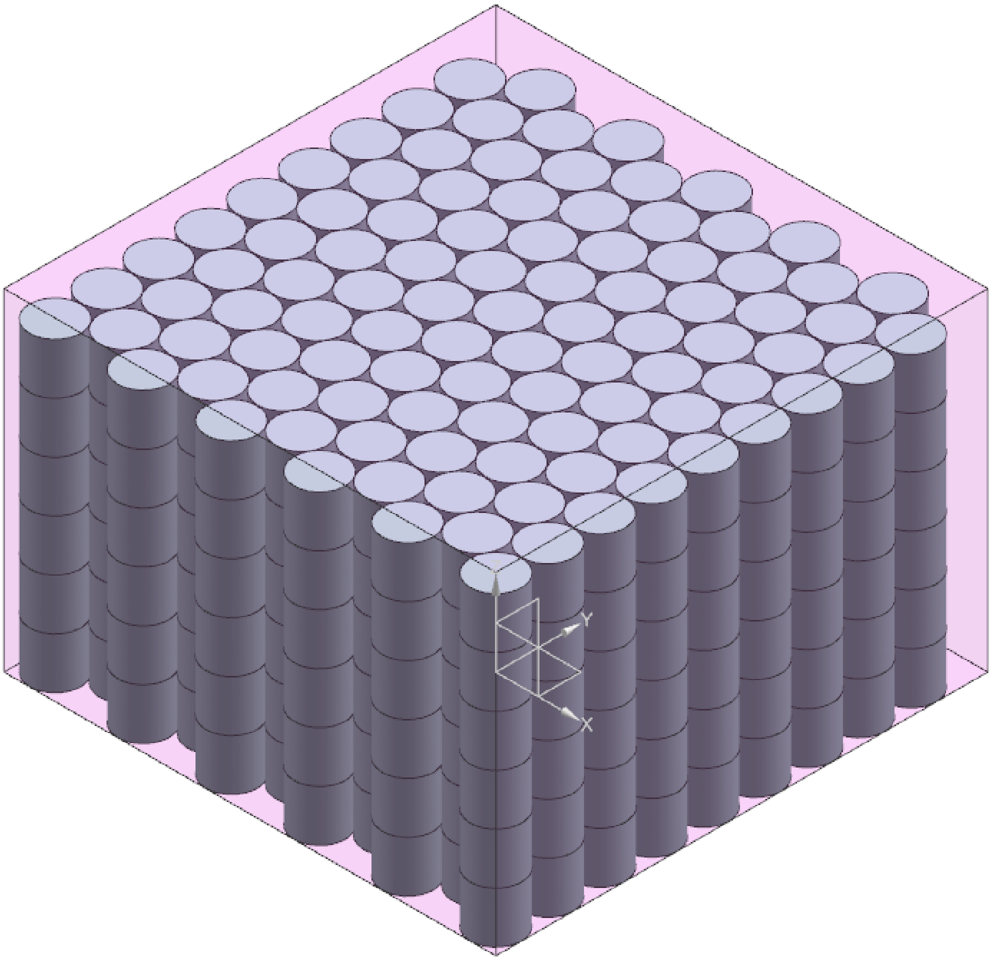


This *almost* uses hexagonal packing, however the circles do not touch the circle above or below them, so it is effectively hexagonal packing which has been reduced in width and increase in height.

Revised Solution

This will give 99 cans per layer, and $\left\lfloor \frac{560}{86.0214} \right\rfloor = 6$ cans stacked vertically, resulting in 594 total cans and a total oil storage volume of $297\ litres$.

This is validated by modelling the cans in 3D, using the optimal arrangement points data, and confirming that they all fit without overlap.



Final Answers

Index	Property	Value (mm or units)
i	Optimal Radius	43.0127
ii	Optimal Height	86.0214
iii	Number of Cans	594

Real World Optimisation

In reality, a balance between material usage and shipping costs would likely need to be made. For example, monetary cost could be used as a measure, and the optimum balance between low material usage and low shipping usage could be found based on current shipping and raw material / processing prices. Alternatively, other costs such as environmental impact could be considered.

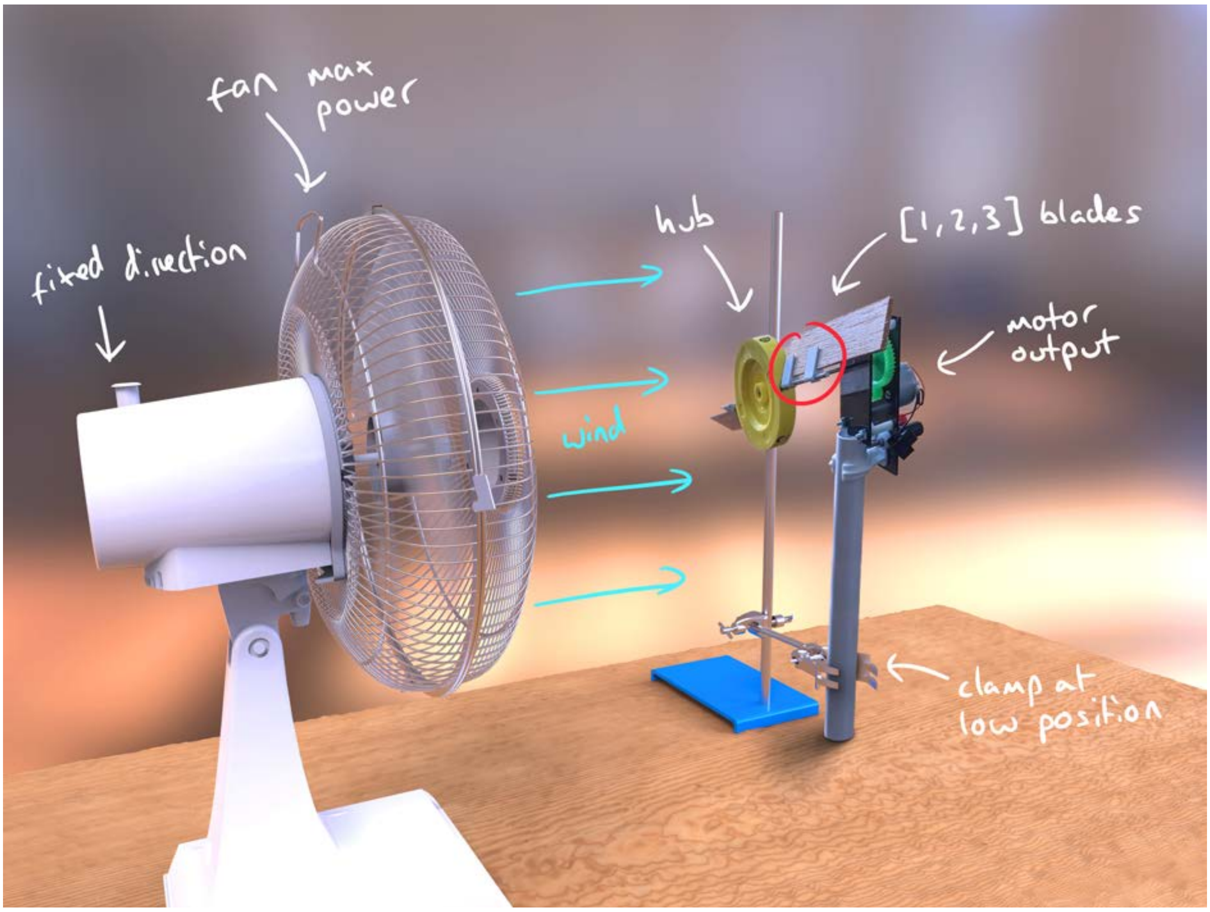
Assignment 1b

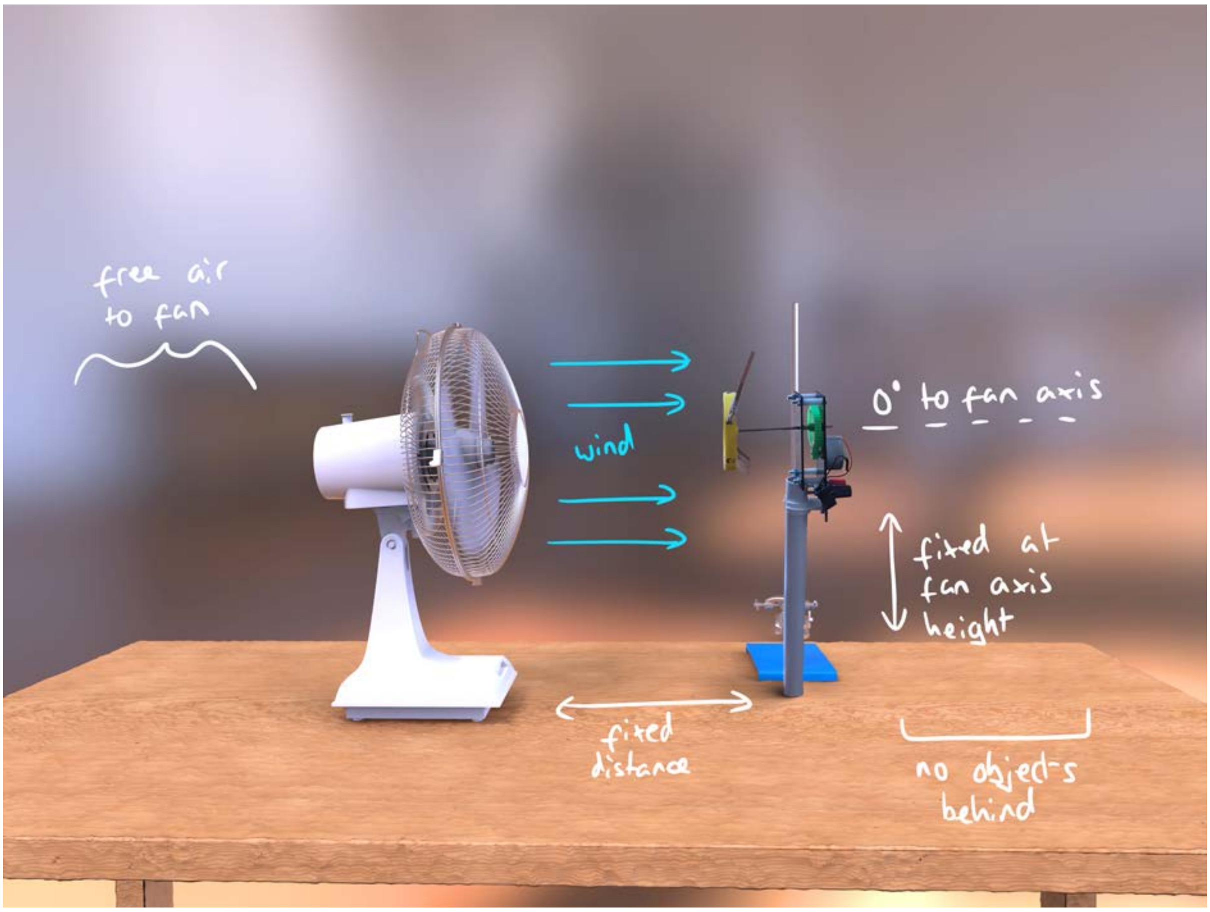
i - Aim & Objectives

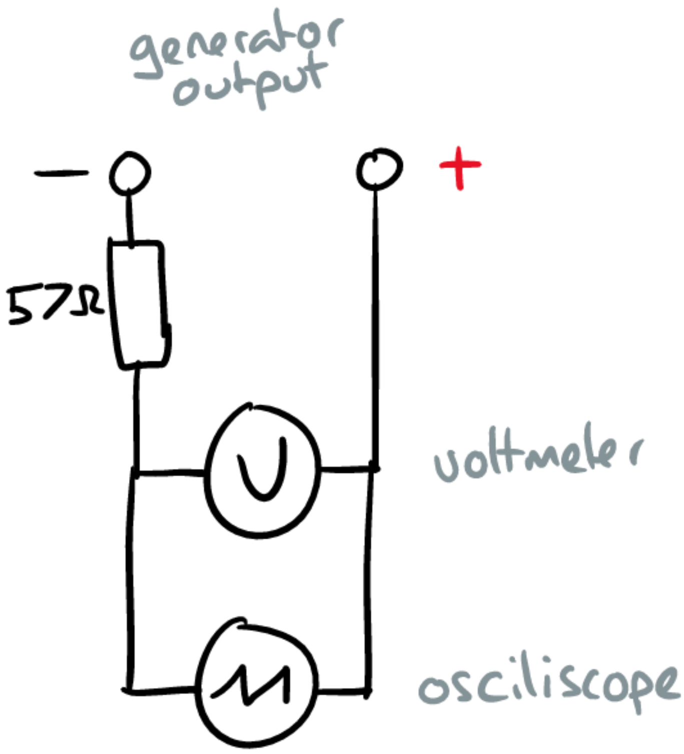
- Confidently determine the interaction between three selected input factors, and an output factor of the small scale wind turbine.
- Select three input factors which can provide useful information when designing and building a wind turbine.
- Conduct practical experiments according to best practices; minimising errors and ensuring valid data.
- Sufficiently minimise the resource (time / material) requirements, without compromising the output data.
- Conduct an analysis of variance, and understand the importance behind statistical significance, and the resultant terms including F-ratio.

ii - Experimental Setup

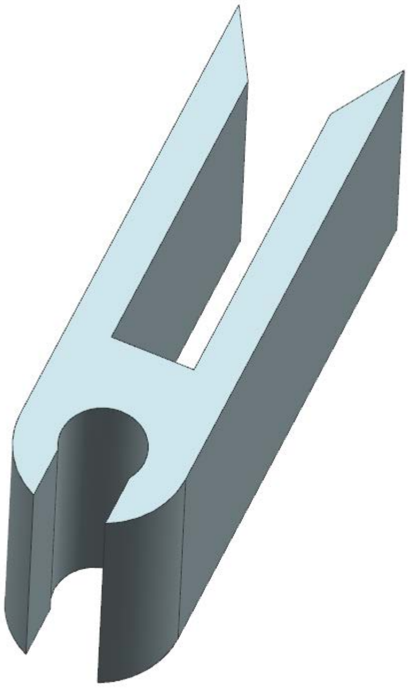
Below are two annotated renders of the experimental setup, and the circuit diagram of the generator output and measurement tools.

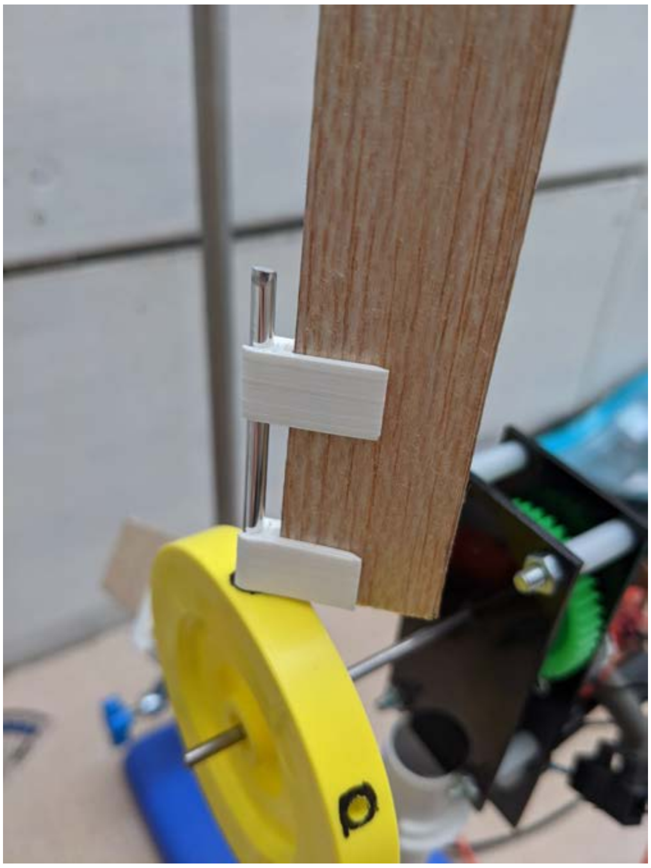

fan max power
fixed direction
hub
[1,2,3] blades
motor output
wind
clamp at low position


free air to fan
wind
0° to fan axis
fixed at fan axis height
fixed distance
no objects behind



To help repeatability, and to minimise rotor movement during the experiment, clips were 3D printed to hold the rotors to the rotor shafts. This removed random errors from tape or glue which could influence the aerodynamic or structural properties of the turbine. Two iterations minimised the size of each clip, and they were a tight interference fit ensuring that no movement (e.g. changing pitch angle) could occur during the test.





Throughout the experiment there were uncontrolled factors which could influence the results. This includes rapidly changing temperatures and external wind / drafts, as the experiments were conducted in an enclosed lean-to.

Further Notes

- To help reduce friction as an uncontrolled factor (e.g. friction changes as the gears heat and cool), the gear train was oiled using 3 in 1 general purpose oil.
- The turbine was allowed to reach equilibrium speed by allowing it to run for ~30s before measurements were taken.
- To ensure the most accurate readings, the oscilloscope was used to log DC average voltage, and frequency of the voltage signal (which is proportional to the rpm). These averages were over a time range of 30s.
- To back-up the oscilloscope data, a multimeter on voltage would also measure the DC average voltage, and a visual check to ensure the values were similar was performed, however only oscilloscope values were recorded to ensure no equipment error was introduced.

iii - Factors & Levels

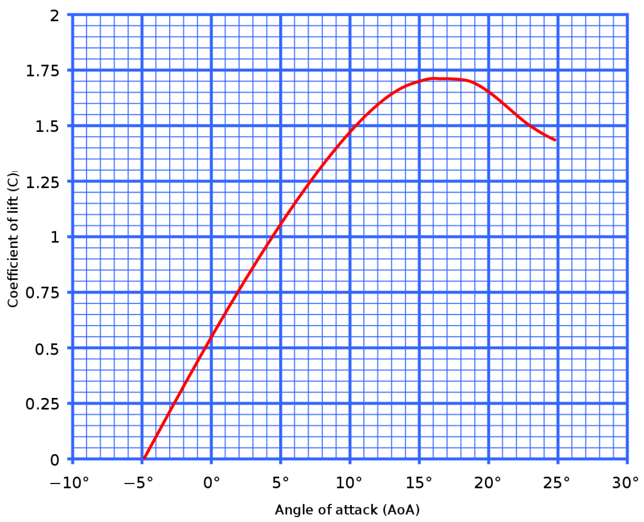
As assignment 1c was conducted before this, the research in that assignment will influence this one.

- Pitch Angle: easy to implement in the practical experiment, and it's relatively clear influence over output power (which should react similarly to rotor rpm) in the tests conducted for 1c.

  Due to the complexity of accurately measuring the wind speed from the fan, it is difficult to equate pitch angle to angle of attack in this small scale experiment. As the wind speed will be approximately constant during the experiment, it can be assumed that $AoA \propto Pitch$. Therefore, data from angle of attack can be used in place of pitch angle in this case.

  For small angles, the coefficient of lift, and resultantly force, and rpm are approximately linear, therefore allowing only two measurements to be taken.

  To attempt to ensure the angle of attack stays small, the pitch angles were selected to be as small as possible while still overcoming friction in the worst case scenario, and ensuring the range between chosen angles was large enough to minimise random errors (for example visual angle measurement errors)

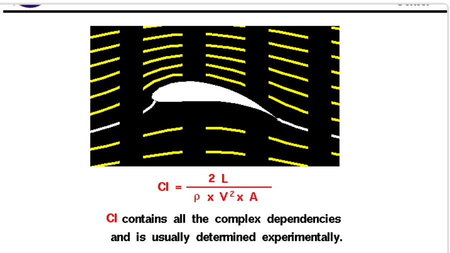
  Approximations are made by assuming the cuboid rotors act in a similar manner to an aerofoil profile.

- Chord length of each blade: also easy to implement. The lift is proportional to the wing area, which will be varied linearly with chord length if span is kept constant. Therefore, only two measurements are needed. To help reduce influence of errors, a large jump in chord length will ensure the largest change in output data. The two chord length values used are 25mm +- 0.25mm, and 75mm +- 0.25mm.

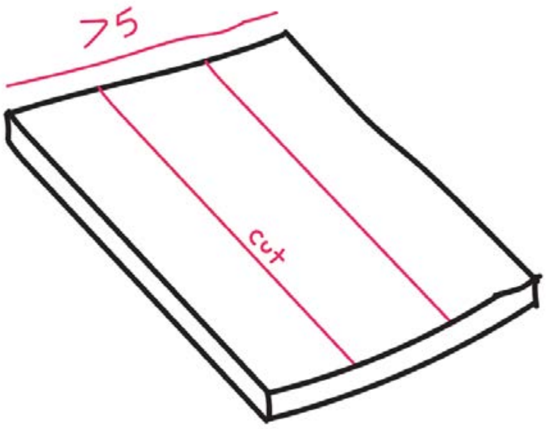
The Lift Coefficient

The lift coefficient is a number that aerodynamicists use to model all of the complex dependencies of shape, inclination, and some flow conditions on lift. This

https://www.grc.nasa.gov/WWW/K-12/airplane/liftco.html



These values were chosen specifically to help minimise material usage. One sheet could be cut into three for chord length C0, and three whole sheets could be used for chord length C1. These whole sheets were therefore undamaged to allow reuse or adaptation in the future.



- Number of blades: More blades will decrease aerodynamic efficiency by increasing the drag (e.g. each blade has a fixed drag force, doubling the blades will ~2x the total drag force).

  This is shown by the drag equation:

$$F_D = \frac{1}{2}\rho v^2 C_D A$$

Doubling the number of blades effectively doubles the area, A. Drag is proportional to the square of wind speed, so it can be assumed that wind speed to drag is not linear. Therefore, three measurements will be taken. 1, 2, and 3 blades will be used as this is possible without modification of the central hub, and ensures constant spacing between levels.

The number of blades will affect the torque produced, which may help overcome mechanical resistance, and generated magnetic fields when current is flowing in the generator. The number of blades also affects the mechanical response of the turbine; two blades can produce oscillatory vibrations as the 'blade plane' goes in and out of alignment with the direction of the wind, resultantly varying the force produced.

When testing with one blade, a second will be positioned on the opposite side, but with a 0° pitch angle to contribute minimally to the generated force. This is less preferred over setting the angle of attack to 0°, but as previously discussed the AoA is difficult to accurately determine given the experimental setup.

This allows the blade to be balanced, and ensures gravity does not influence the speed of the turbine more than it would with two or three blades. This does have an effect on the drag of the turbine, which may influence the results, however the unbalanced alternative would likely produce less favourable results.

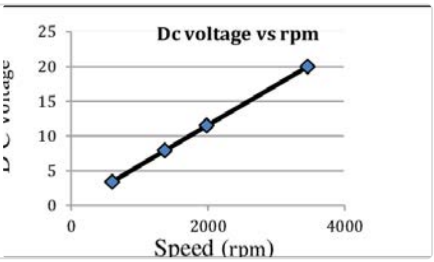
Other factors

- Wind speed would have been ideal to test, however:
    - *Voltage output is proportional to rpm, and voltage is proportional to power ($P = VI$).*

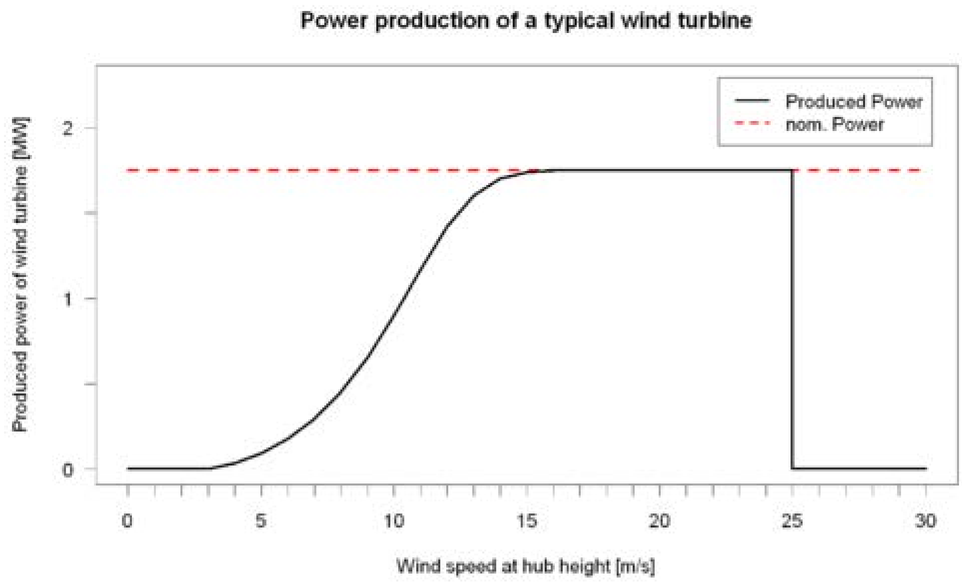
Fig 11: Graph of DC Voltage v/s Speed in RPM

Download scientific diagram | Graph of DC Voltage v/s Speed in RPM from publication: Design and Development of Solar Hybrid Bicycle | Since the fuel

https://www.researchgate.net/figure/Graph-of-DC-Voltage-v-s-Speed-in-RPM_fig7_305683710



-
    - *The power output (and resultantly RPM) of a turbine is **not** proportional to wind speed.*



  - *Therefore, more than two measurements would need to be taken. With the supplied fan, this is not possible.*
- Control factors such as yaw angle (angle between wind direction and wind turbine direction) would be interesting to test, however real-world turbines are designed to minimise this and therefore data obtained may not be extensively useful.

iv - Experimental Dataset

Refer to `data_cleaned.csv`.

Experiments are coded according to below table.

```
ID Key
#    - Repeat number     ( 0:One  |  1:Two  |  2:Three  |  3:Four )
 #   - Number of blades  ( 0:One  |  1:Two  |  2:Three )
  #  - Pitch angle       ( 0:20d  |  1:30d )
   # - Chord length      ( 0:25mm |  1:75mm )
```

v - Results

Source of Variance

Source	SSQ	DoF	Mean SQ	F Ratio	P Values	Significant at 0.05

Source	SSQ	DoF	Mean SQ	F Ratio	P Values	Significant at 0.05
P	4.549161	1	4.549161	29.56602	<0.00001	☑
B	28.04891	2	14.02445	91.14809	<0.00001	☑
C	55.51516	1	55.51516	360.8055	<0.00001	☑
PB	0.307032	2	0.153516	0.997736	0.375421	☐
PC	0.796448	1	0.796448	5.176293	0.026892	☑
BC	11.15876	2	5.579382	36.26166	<0.00001	☑
PBC	1.438866	2	0.719433	4.675758	0.013403	☑
Within, Y	5.539122	36	0.153864			☐
Total, T	107.3535	47	2.284116			☐

vi - Significance

All of the results show statistical significance at a significance level of 0.05, except for the source from interaction between P (pitch angle) and B (the number of blades). Therefore, for the significant sources, we can reject H_0 (the null hypothesis that the input variance of the sources has no effect on the output variance of the turbine voltage), in favour of H_1, suggesting that the input variance of the sources does have an effect on the output variance of the turbine voltage.

vii - Discussion

The F ratio is an indication of how impactful the input factor is on the output factor. It can be described as a ratio of the F value which describes the variance obtained through the experimental data, and an F ratio which describes the variance of the data if H_0 is true, i.e. there is no relationship between the input and output factors, and any variance comes from randomness in the experiment and measurements.

If in reality, H_0 is true, then the numerator and denominator are approximately equal, and so the F ratio is approximately 1. However, if the numerator is much greater than the denominator, it indicates that the amount of variance is much greater than can be explained by randomness in the experiment - and must be explained by something else; that there is a relationship between the input and output factors.

Based on this, we can see that the source of chord length variation has the highest F ratio, and therefore causes the highest amount of variance in the output data which cannot be explained by randomness.

Degrees of freedom does factor into the significance of this result, however this specific source has a DoF of 1, and therefore can be said to have the highest F-ratio regardless of the other sources DoF.

This result is not surprising as the chord length is a defining factor in the coefficient of lift equation (and resultantly a defining factor in influencing the lift of the aerofoil), in the form $S = cs$ (area = chord length * span)

$$C_L = \frac{L}{\frac{1}{2}\rho u^2 S}$$

This makes sense when considering the practical effect of chord length. If lift is created by a difference in pressure above and below the wing, and force is pressure multiplied by the area on which it is acting, having a longer chord allows the pressure to act on a greater area, increasing force.

The same argument can be made when considering the effect of number of blades B on the output voltage, when looking at the blade element expression for change in force (for a rotor element).

$$dF_x = \frac{B\rho c {v_1}^2 (1-a)^2}{2sin^2\theta} \times C_x \; dr$$

However, existing research revealed that increasing the number of blades indefinitely is not the optimal solution. Possible reasons for this is that adding more blades will result in turbulent flow onto the blades, as caused by other blades on the turbine, which allows for less energy extraction than ideal laminar flow.

In reality, increased number of blades will increase the maximum possible torque on the rotors, which is used to overcome friction within the system. The small scale apparatus likely has much more friction proportionally than a real wind turbine (even though the gear train was lubricated before testing), and therefore the increased torque availability had a much higher impact than in a full-scale turbine.

Although the above discussion refers to force and not the output voltage, the two are related through output rpm. In an ideal generator, output voltage is a linear function of rpm. The rpm during the experiment is at equilibrium; the point at which all system forces are balanced. This means the lift force from the rotors will balance the friction forces, which is also a function of rpm. Therefore, it is suitable to refer to force on the rotor when discussing output generator voltage.

The only insignificant factor was the product of number of blades and pitch angle. Therefore, there is no interaction between these two factors which results in a significant effect on the output voltage variance.

Although the two sources had significant effect on the output variance when taken in isolation, a lack of significance of their product suggests that varying the number of blades had a similar effect on output variance, regardless of which pitch angle it was at. Similarly, varying the pitch angle had a similar effect on output variance, regardless of the number of blades.

An overview of experimental limitations which could have impacted the results are shown in the table below.

Experimental Limitation

Name	Description
Friction in small scale turbine	Likely much more impactful proportionally than a full scale turbine. May vary over the course of the experiment as lubricant disperses.
Environmental temperature variation	Tests were conducted in the middle of the day when temperature should be approximately constant, and randomisation of experiments should help reduce impact. Lower temperature may result in denser air and higher force for given controlled factors.
Environmental wind variation	Tests were conducted in an enclosed lean-to with noticeable drafts, which could randomly effect the output voltage.
Accuracy of rotor blades	Hand cut from fairly uncontrolled balsa wood. Variations in dimensions could unintentionally vary chord length or span, or blade profile. Density variations could vary dynamic response of the system.
Turbine vibrations	At high rpm, the turbine would vibrate and flex much more than at lower rpm. Therefore, energy extracted by the blades could be converted into vibration instead of generated power, which could reduce the voltage by a greater amount for higher rpm tests.
Visual pitch angle measurements	The pitch angle was measured with a protractor by eye. This results in a maximum precision of ~0.5° however difficulties with visual measurements result in an accuracy of likely much less.

Name	Description
Manual experiment setup	As blades were frequently removed and attached to the experiment, their positions (such as distance from the hub) may vary slightly between tests. Although measures were taken to minimise this, such as printed clips, this could still influence the results.

Assignment 1c

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

Aa 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.

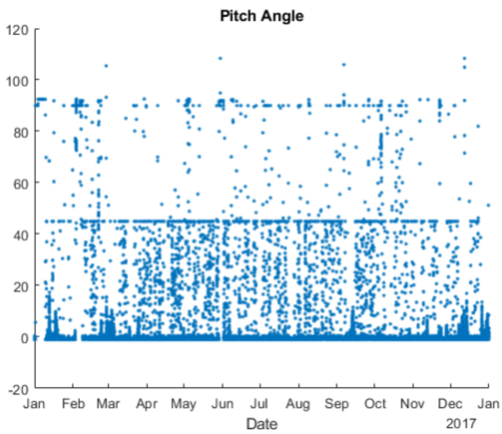
```
% Remove any rows which contain missing values
data = rmmissing(data)
```

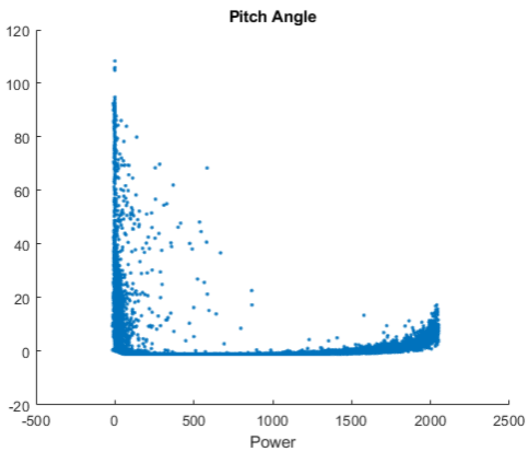
3. The `Date_time` string column is converted to a `datetime` object which can be used to plot in chronological order.

```
% 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





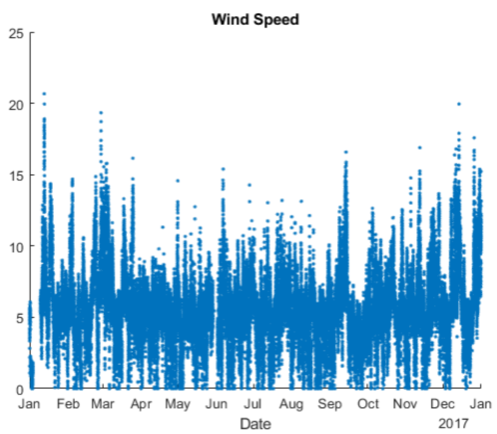
Some features immediately stand out - there is a visible line at 45°, suggesting this angle is used commonly, however there is no obvious correlation between 45° and high power, indicating this may be a default value when the turbine is not in use. The same can be said for 0°.

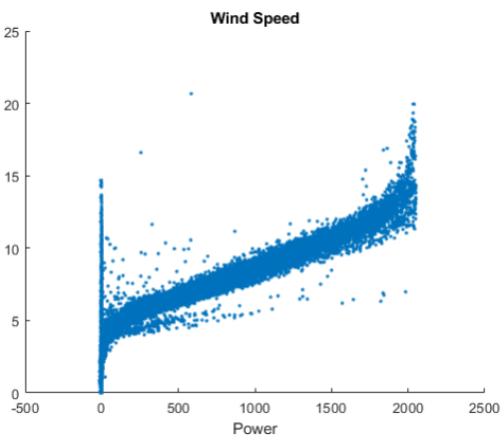
Few values appear above 90° - although specific information about the turbine could not be found through research, it can be assumed that the blades are not supposed to rotate past 90°, and therefore that these values are outliers. These values are removed.

```
% 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



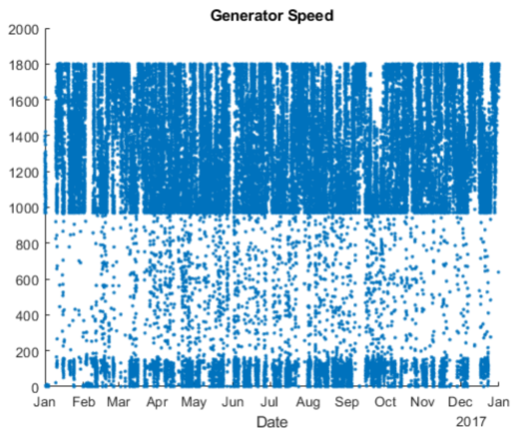


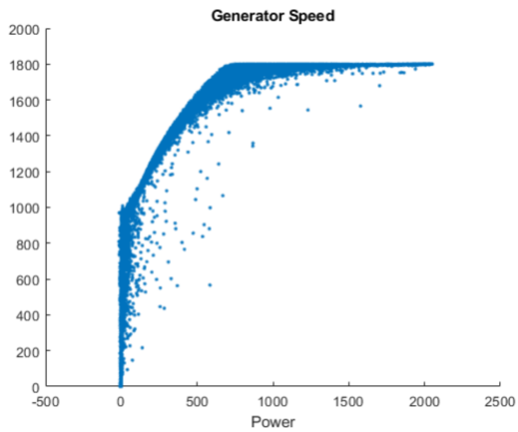
Removing data points that are a result of the control system, rather than the turbine physics, are recommended. It can be seen that there are a lot of data points where wind speed is $\neq 0$, however power is 0. This could be indicative of the turbine being stopped when electricity demand is already 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.

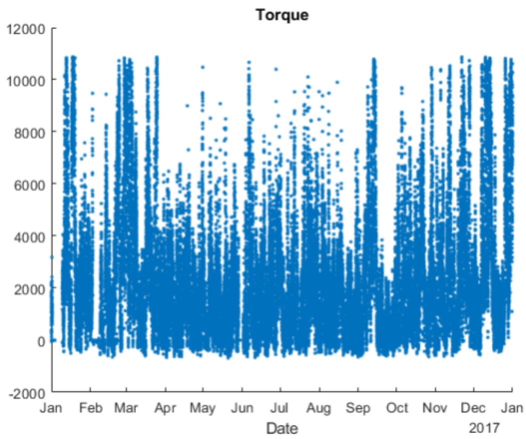
Generator Speed

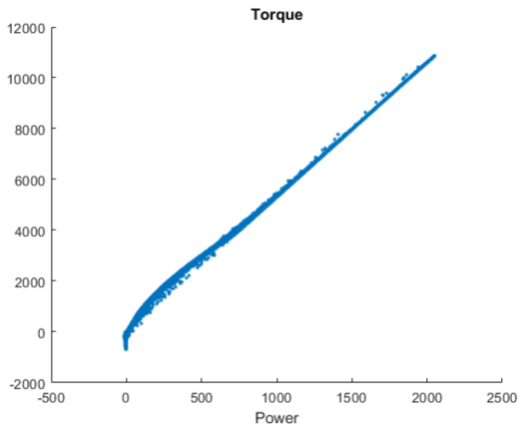




It could be argued that generator speeds below 1000 are as a result of system control factors - however there is not enough information known about the turbine to justify this, and therefore all the data will be kept.

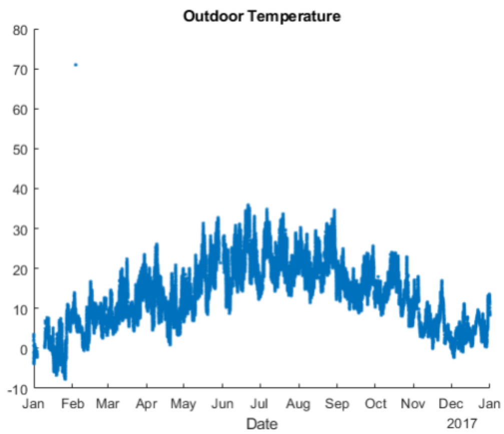
Torque

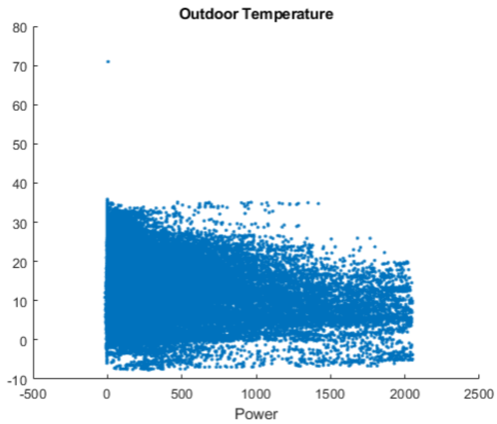




Appears to be a very obvious positive correlation, with no noticeable outliers. Therefore, the data will not be changed.

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.

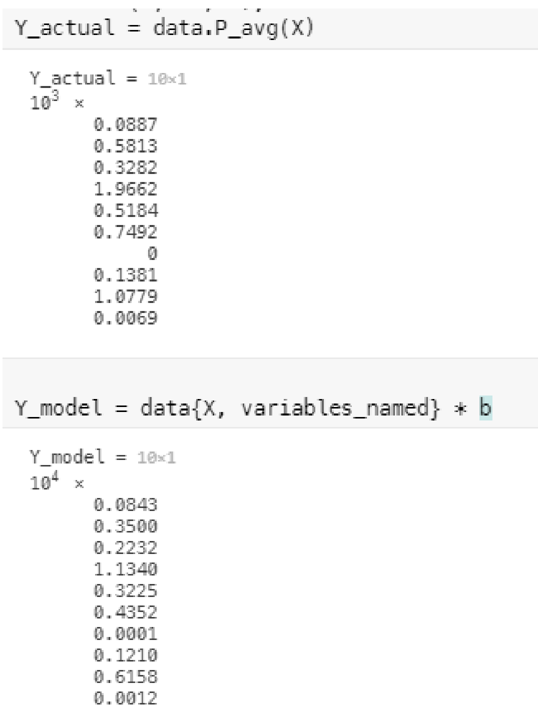
Meanwhile, the other factors have little influence over the output power variance according to the values obtained. This could be indicative of the output factor being less *sensitive* to these input factors; that they need to be varied a greater amount (in terms of standard deviation) than the torque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

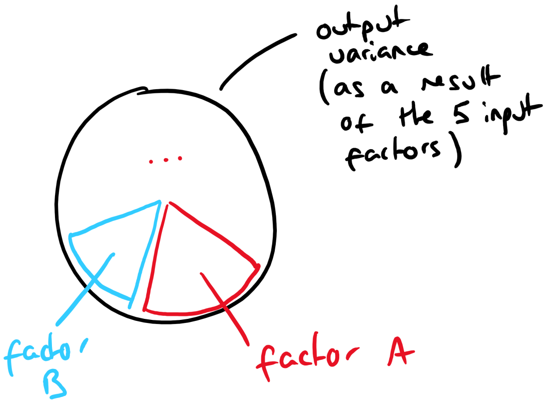
- Due to the small differences between S_i and S_{Ti}, this suggests the factors are fairly independent, and interactions between them do not contribute to the variance of the output.
- Wind speed is the most significant factor when compared to the other tested factors, for influencing the output variance.
- 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.

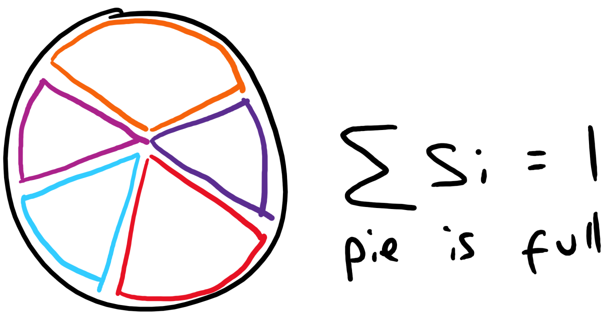
Notes on the sum of S_i and S_{Ti}

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.

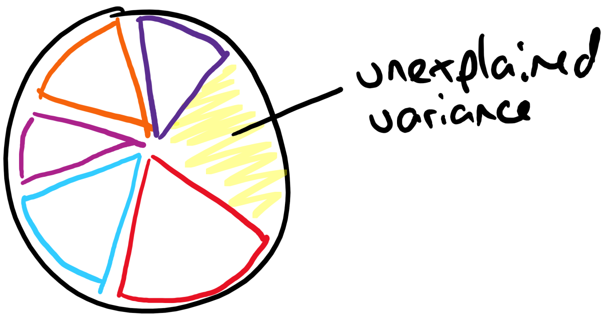
The S_i or S_{Ti} values represent the proportion of the pie that that factors slice occupies. In other words, a value of 0.5 means that input factor can explain half of the output factor variance.



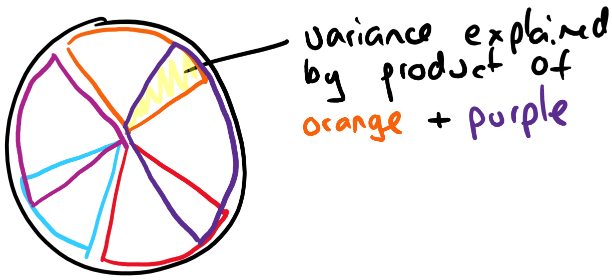
If $\sum S_i$ or $\sum S_{Ti} = 1$, this means that the variance of the input factors perfectly explain the variance of the output factor. They each contribute some proportion to the total output variance.



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.

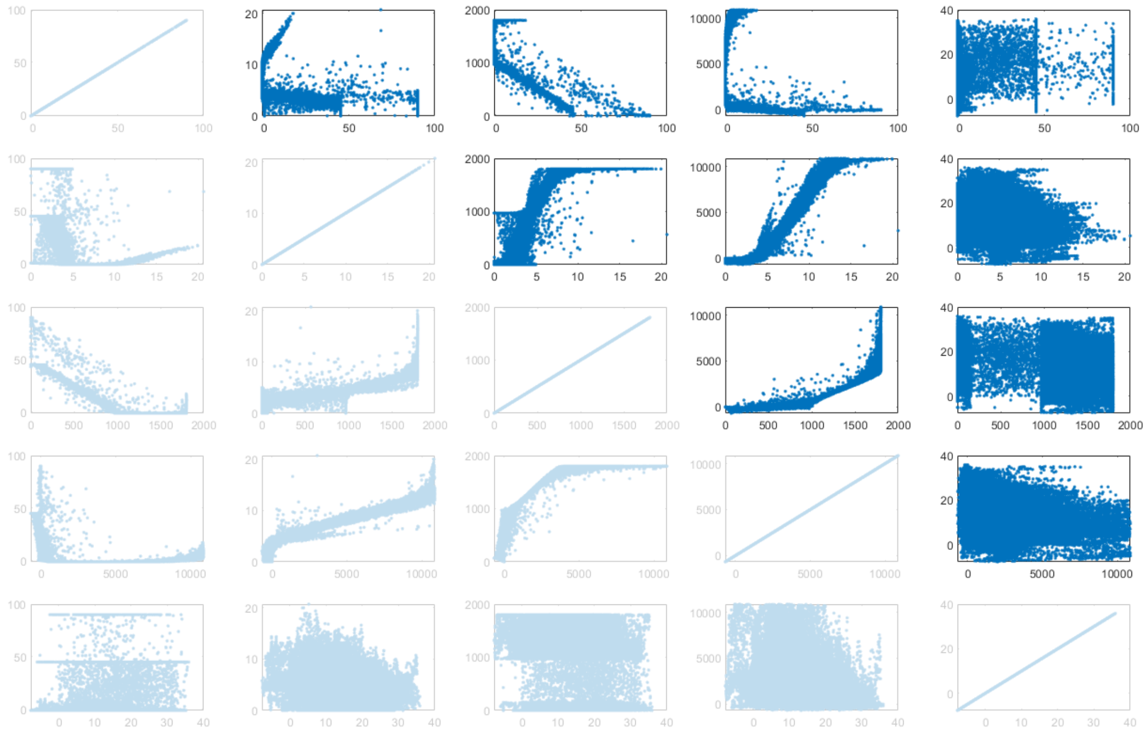


It can be seen that this overlap of explained variance should equal the missing section of the S_i pie. Mathematically, this means that $S_i + S_{Ti} = 2$, as together they create two whole pies.

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.

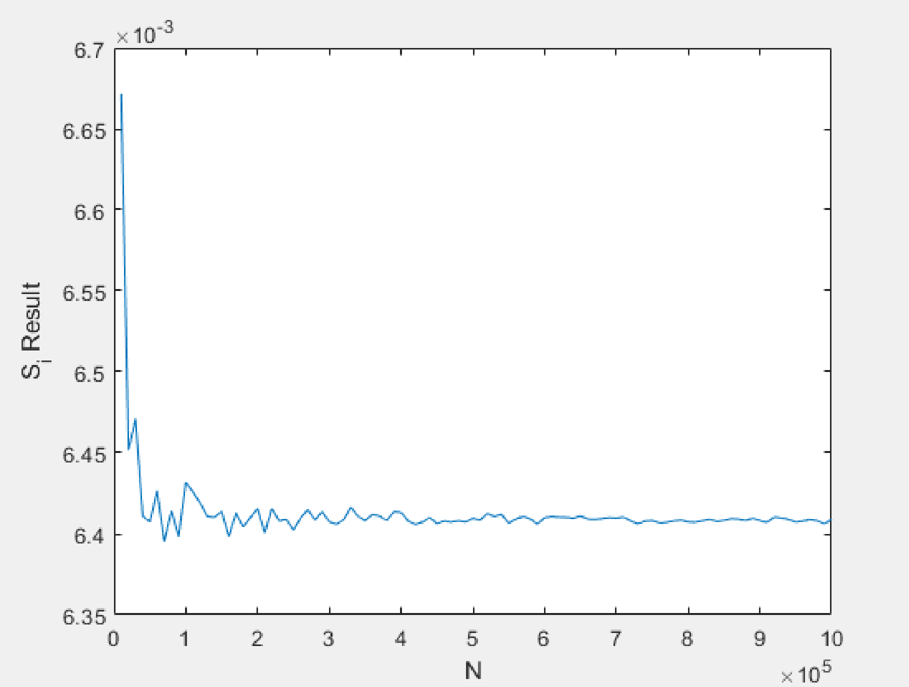


Visually inspecting the plots can help determine if there is a strong correlation between any two input factors. There is some correlation between some input factors, most notably at element (2, 4) (Wind Speed and Torque), which may promote caution regarding results obtained for these factors, however 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.