Sensing Technologies and Mathematics for Geomatics
GEO1001.2020
MSc Geomatics
Delft University of Technology

Assignment 1

Theodoros Papakostas (5287928)

Introduction: For this assignment we use data collected from 5 heat stress sensors placed somewhere in the Netherlands during this summer. The sensors are Kestrel 5400 and their specs are included within the assignment materials. In order to identify if the dataset is of any value to our "employer", it is our job to deeply analyse the dataset and derive hypothesis from it.[1]

1 Lecture A1

1.1 Mean Statistics

First of all, given the datasets of 5 sensors(A,B,C,D,E) with measurements over 19 variables, mean statistics (mean, variance and standard deviation) were computed.[1]The results are aggregated on the matrices underneath this subsection.

In general, the mean statistics(mean,variance,standard deviation), fluctuate in ordinary values and no peculiar behaviors over them can be identified. Thus, we cannot yet reach any statistically significant conclusions over them, nor derive any robust hypothesis from them. Especially when, we have to do with 19 different variables, leads to the need of more circumstantial analysis.

Sensors	Statistics	Direction β True	Wind Speed	Crosswind Speed	Headwind Speed	Temperature	Globe Temperature	Wind Chill	Relative Humidity	Heat Stress Index	Dew Point	Psychro Wet Bulb Temperature	Station Pressure	Barometric Pressure	Altitude	Density Altitude	NA Wet Bulb Temperature	WBGT	TWL	Direction β Mag
SENSOR A	Mean	209.41	1.29	0.96	0.16	17.97	21.54	17.84	78.18	17.9	13.55	15.27	1016.17	1016.13	-25.99	137.32	15.98	17.25	301.39	208.91
	Variance	10108.9	1.25	0.93	1.03	15.86	68.19	16.26	376.01	15	9.72	6.94	38.47	38.47	2663.64	26510	10.01	16.14	814.77	10105.7
	Standard deviation	100.54	1.12	0.96	1.02	3.98	8.26	4.03	19.39	3.87	3.12	2.64	6.2	6.2	51.61	162.82	3.16	4.02	28.54	100.53
SENSOR B	Mean	183.41	1.24	0.84	-0.13	18.07	21.8	17.95	77.88	18	13.53	15.3	1016.66	1016.62	-30.06	135.58	16	17.32	299.45	183.22
	Variance	9977.22	1.3	0.88	1.26	16.63	66.05	17.04	408.62	15.44	9.64	6.77	36.84	36.83	2545.71	26863.3	9.81	15.84	790.07	9975.45
	Standard deviation	99.89	1.14	0.94	1.12	4.08	8.13	4.13	20.21	3.93	3.1	2.6	6.07	6.07	50.46	163.9	3.13	3.98	28.11	99.88
SENSOR C	Mean	183.59	1.37	0.96	-0.26	17.91	21.59	17.77	77.96	17.83	13.46	15.2	1016.69	1016.65	-30.34	129.62	15.93	17.23	301.9	183.08
	Variance	7703.36	1.43	1.04	1.27	16.1	67.94	16.54	374.62	15.36	10.08	7.24	37.69	37.68	2608.53	26986.6	10.48	16.55	766.53	7704.62
	Standard deviation	87.77	1.2	1.02	1.13	4.01	8.24	4.07	19.36	3.92	3.18	2.69	6.14	6.14	51.07	164.28	3.24	4.07	27.69	87.78
SENSOR D	Mean	198.33	1.58	1.21	-0.3	18	21.36	17.84	77.94	17.92	13.51	15.26	1016.73	1016.69	-30.65	132.41	15.92	17.18	305.25	197.83
	Variance	8133.89	1.74	1.45	1.23	16.11	61.2	16.56	389.86	15.12	10.07	7.04	34.99	34.95	2419.72	26516.1	9.99	15.51	616.01	8135.32
	Standard deviation	90.19	1.32	1.2	1.11	4.01	7.82	4.07	19.74	3.89	3.17	2.65	5.92	5.91	49.19	162.84	3.16	3.94	24.82	90.2
SENSOR E	Mean	223.96	0.6	0.44	0.19	18.35	21.18	18.29	76.79	18.29	13.56	15.41	1016.17	1016.13	-25.96	150.84	15.94	17.19	284.12	223.9
	Variance	9308.29	0.51	0.32	0.32	19.04	63.22	19.14	406.49	18.48	9.42	7	38.94	38.94	2692.35	29714.9	9.43	15.49	1289.91	9268.01
	Standard deviation	96.48	0.72	0.56	0.56	4.36	7.95	4.37	20.16	4.3	3.07	2.65	6.24	6.24	51.89	172.38	3.07	3.94	35.92	96.27

Figure 1: Mean statistics of the 5 sensors

1.2 Histograms for the 5 sensors values

Given the Temperature values of the 5 sensors, 2 histograms were created: 1 using 5 bins and 1 using 50 bins.

Those histograms graphically display the shape of the distribution, for each sensor's measurement values.This is really useful in particular, because we have to deal with a big load of observations. However, in order to take full advantage of this specific visualisation, the choice of bin number to be used is very important. Rice's formula on the choice of bin number, is generally used :

$$2 \times \sqrt[3]{N}$$

In our case, the two different bin choices we made, result in two apparently different histograms. On the on side, as we observe on Figure 2, the 5-bin selection leads to a very generic visualisation of the distribution which apparently disregards any peculiar data fluctuation impacts. Thus, even though we have acquired a large number of observations we cannot make full use of it to identify discrete unordinary data behavior,of the 5

sensors. On the other side, as we observe on Figure 3, the 50-bin selection demonstrates a very detailed distribution of the 5 sensors' measurements out of which, useful information can be extracted.However, this accumulated detail, may lead to difficulties in discriminating the correct information. Consequently, the choice of the number of bins to be used is very important.

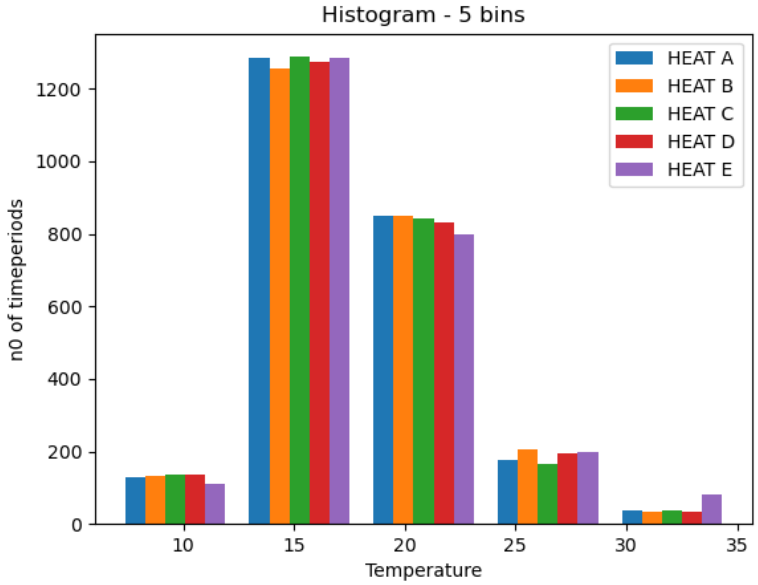

Figure 2: Histogram -5 sensors Temperature

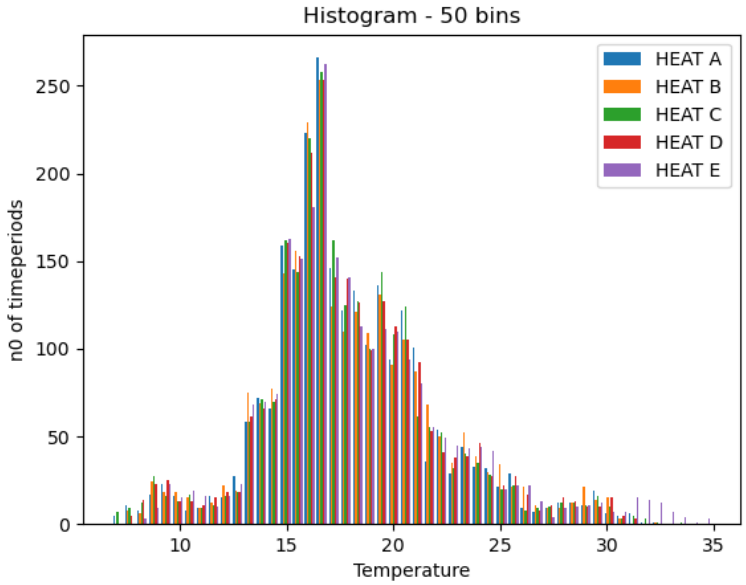

Figure 3: Histogram -5 sensors Temperature

1.3 Frequency polygons for the 5 sensors' Temperature

On this subsection, a plot was created, where frequency polygons for the 5 sensors Temperature overlap in different colors. The No of bins is 27. Choice was made using the Rice's formula for N=2476.

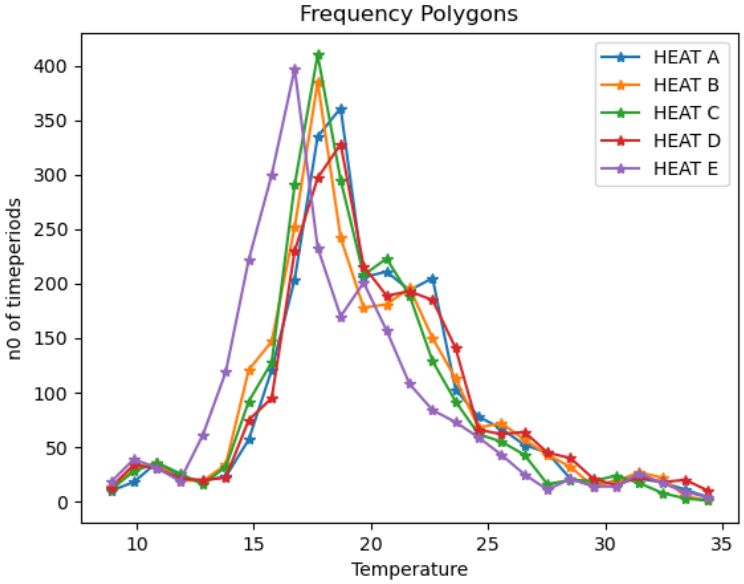

Figure 4: Frequency polygons

1.4 5 sensors boxplots for: Wind Speed, Wind Direction, Temperature

Over here, 3 particular boxplots were generated, containing information for the 5 sensors Wind Speed, Wind Direction and Temperature. The boxplots,in general, are used to identify any outliers and for comparing distributions.

In particular, on the created boxplots, we can observe that for those 3 variables, the 4 sensors behave almost similarly. On the contrary, the E sensor features some dissimilarities concerning its outliers. This is really aparrent, especially in the boxplot for the Wind Direction.

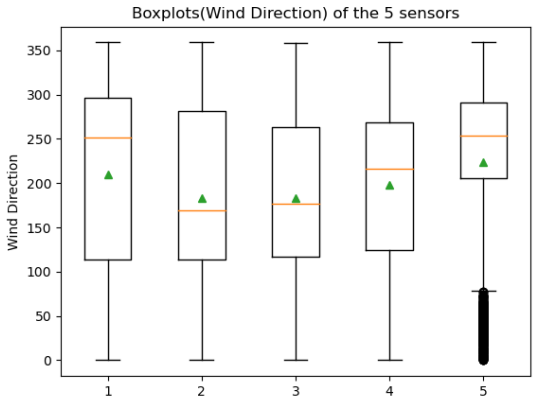

Figure 5: Plot 1

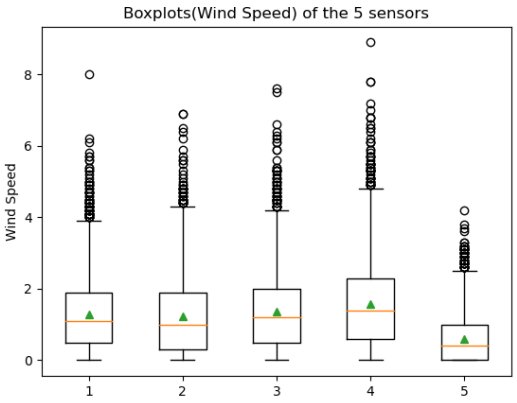

Figure 6: Plot 2

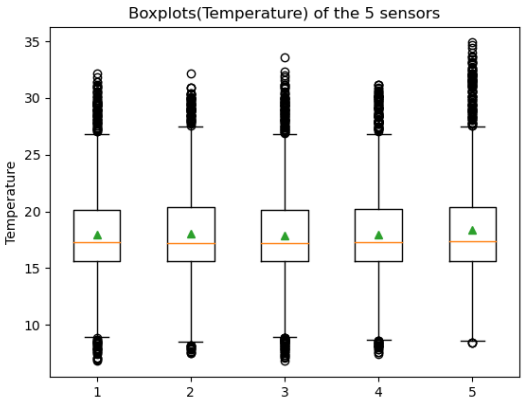

Figure 7: Plot 3

2 Lecture A2

2.1 PMF, PDF, CDF for the 5 sensors' Temperature

Using the 5 sensors' Temperature variable, the PMF,PDF,CDF of each one was plot.

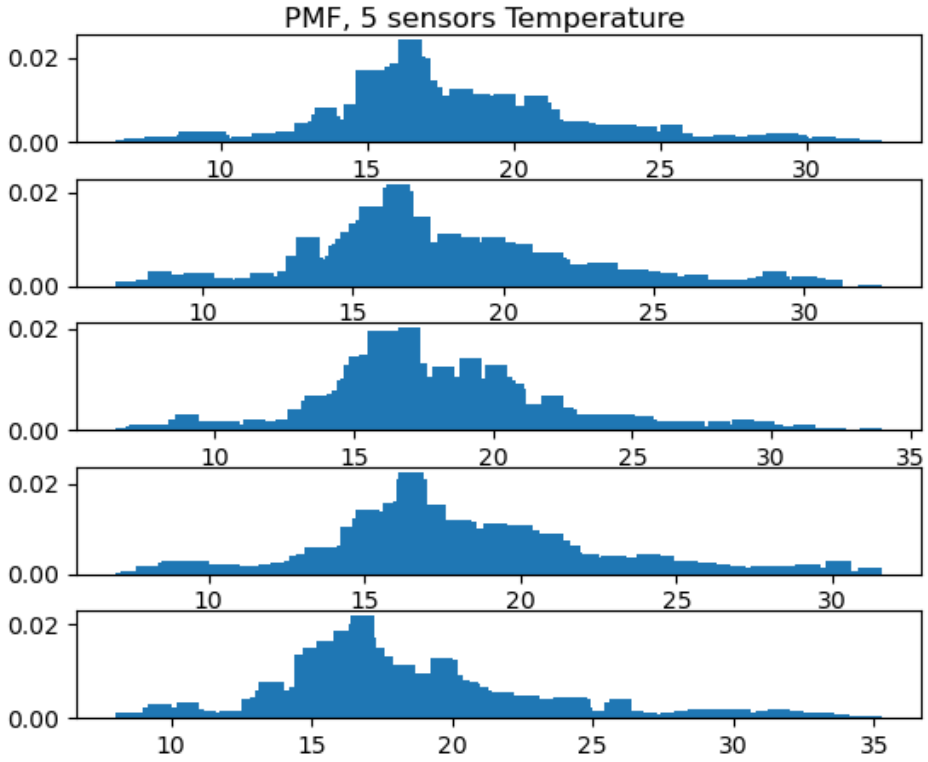

Figure 8: Probability Mass Function, 5 sensors

Probability Mass Functions(PMFs), is a way to represent a distribution of each value's probability. It is a bar chart where we get from frequency to probability, through the division by N(number of values), in order to succeed normalization. A way not to be mislead by the different sample sizes. Indeed, the temperature variable values were not the same for every sensor. In general terms, the PMFs for the 5 sensors, have the same behavior concerning their main bodies. As for their tails, we observe that the probability rates decline proportionally to the maximum and minimum temperature values.

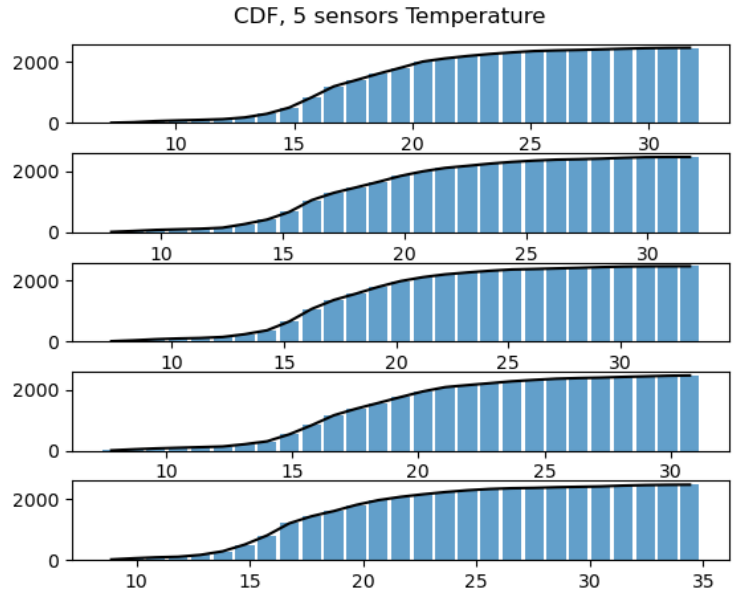

Figure 9: Cumulative Density Function, 5 sensors

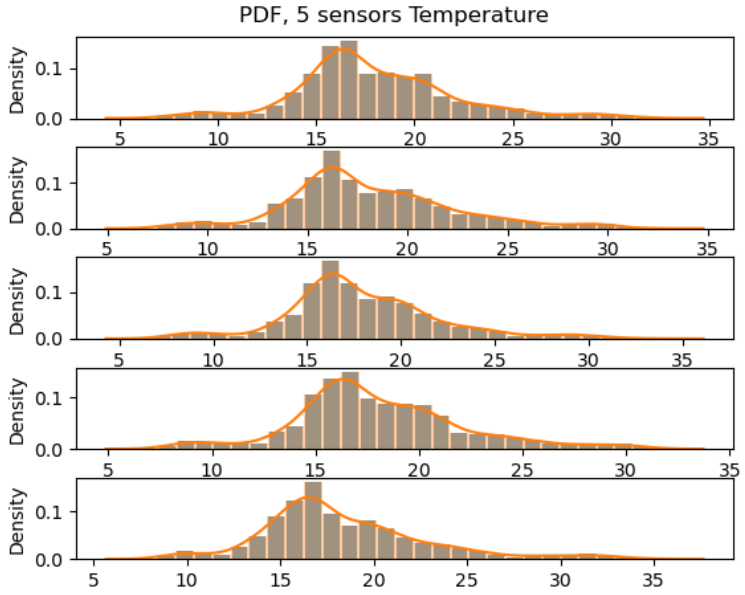

Figure 10: Probability Density Function, 5 sensors

The Cumulative Density Function(CDF) map from a value to its percentile rank. Its derivative is called Probability Density Function(PDF) and actually measures the probability pes unit of x.

The behavior of both distributions CDF and PDF, looks similar in general, for the 5 sensors, and there is not any significant disimilarity to be noted.

2.2 PDF and Kernel density estimation, 5 sensors Wind Speed values

For the Wind Speed values, the PDF and Kernel Density Estimation were plotted.

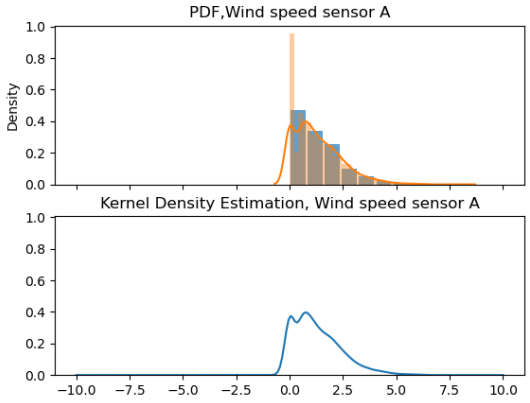

Figure 11: sensor A - Wind speed

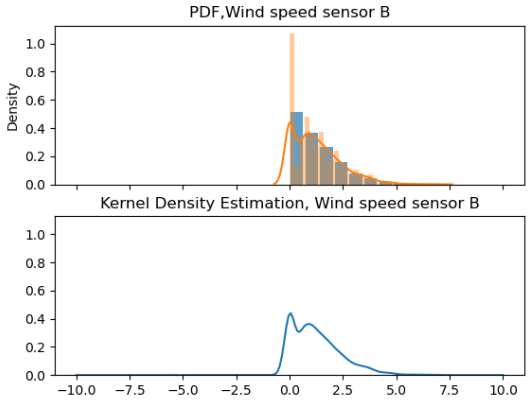

Figure 12: sensor B - Wind speed

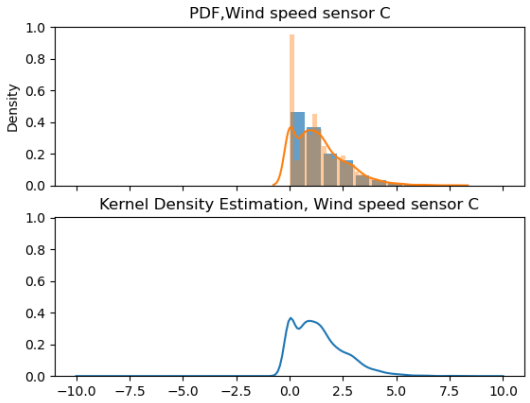

Figure 13: sensor C - Wind speed

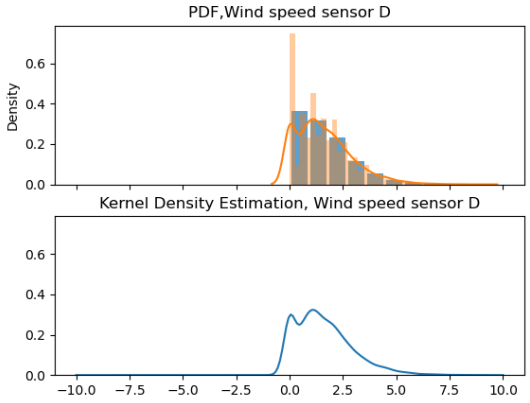

Figure 14: sensor D - Wind speed

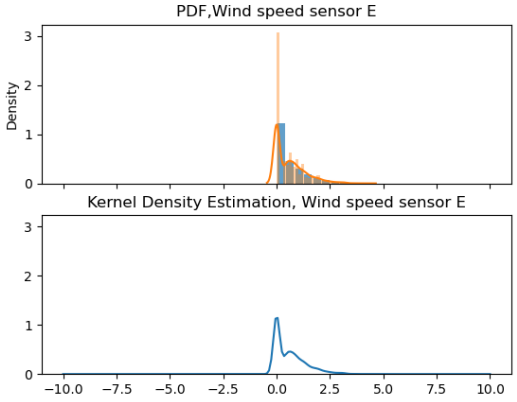

Figure 15: sensor E - Wind speed

We can observe that for all the 5 sensors, the PDF and the KDE are almost identical. Of course, this comes as no surprise, since the KDE is actually an algorithm takes a sample and constructs an appropriately smooth PDF that fits those data. Thus, that's what happens in this incident, with the Wind values of each sensor.

3 Lecture A3

3.1 Coorelation and Coefficients

Correlations were computed between all the sensors for the variables: Temperature, Wet Bulb Globe Temperature (WBGT), Crosswind Speed. For this purpose, Pearson's and Spearmann's rank coefficients were used. At last, a scatter plot with both coefficients with the 3 variables was plot.

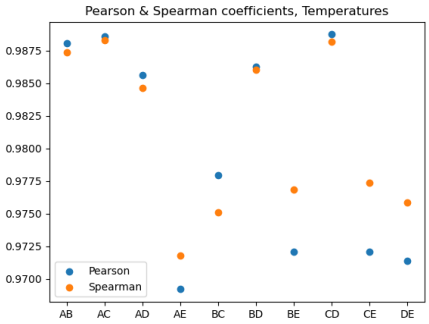

Figure 16: Coorelation - Temperature

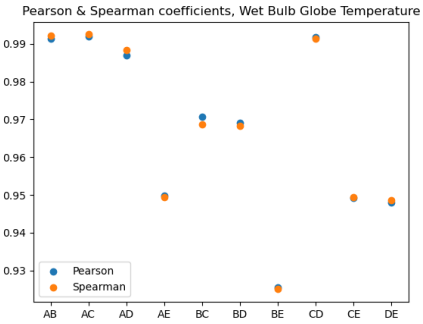

Figure 17: Coorelation - Wet Bulb Globe Temperature

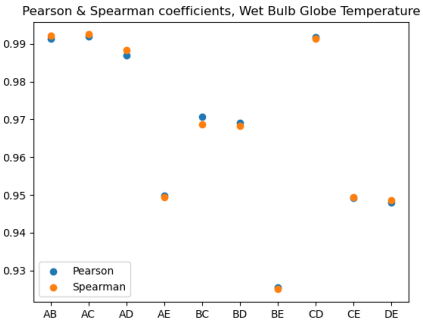

Figure 18: Coorelation - Crosswind Speed

3.2 Coorelations

Something that can easily be spot, on the scatter plots of the coefficients, is that both Temperature and WBGTemperature have high coorelation(values are over 0.97), for all the sensors. This is logical because they both have to do with temperature values which is less likely to have great ups and downs in a particular place. Addittionally, those variables' values of all the sensors have a linear relationship that could be seen from the values scatter plot, which also shows their great coorelation. On the contrary, Crosswind speed has a lot smaller coorelation values between the sensors(from 0.4 to 0.65), due to its non-linear relationship that could be seen from the values scatter plots. Generally we note great coorelation between the sensors' values for Temperature,Wet Bulb Globe Temperature, and a lot less coorelation between the sensors' values for Crosswind speed.

3.3 Hypothesis

As mentioned before, if we observe the scatter plots of the values for the variables Temperature, Wet Bulb Globe Temperature and Crosswind Speed we reach the conclusion that for the 5 sensors the relationships between Temperatures are linear and between Crosswind Speeds are not linear. We take advantage of the Spearman rank coefficients that are more robust for non linear relationships, and thus, we note that in the Crosswind Speed coefficients scatter plot the sensors C and D have the greatest coorelation. In parallel, this high coorelation between C and D is also visible in the other two coefficient scatter plots(temperature). That's why I assume that C and D must be really close, and may be the two neighbour sensors on the left side. Now, I observe that C and D are less coorelated on the Crosswind scatter, with E sensor, which I assume that is the one top right. Also, on the Temperature scatter plots, we can observe that AC,AD have high coorelation values,higher that BC,BD so we can assume the exact position of A and B(B up, A down, because it is closer to C and D).So, at last, we can conclude to the final hypothesis for the sensors' locations.

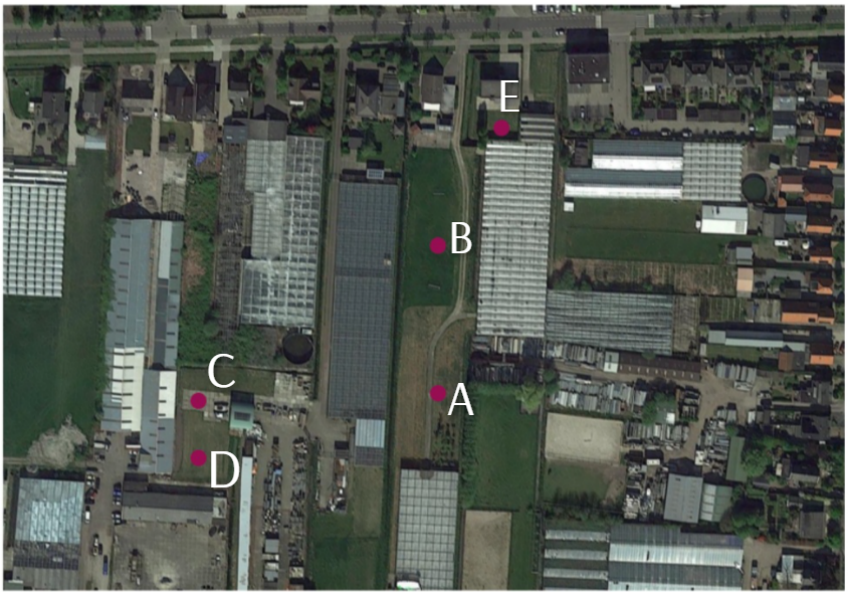

Figure 19: Hypothetical sensors locations

4 Lecture A4

4.1 CDF for 5 sensors' Temperature and Wind Speed, 95 percent confidence intervals

In this part, the CDF for all the sensors and for variables Temperature and Wind Speed, were plot. Then the 95 percent confidence intervals for variables Temperature and Wind Speed for all the sensors were computed and saved in a table (csv form). The table can be seen underneath.

	Confidence Intervals,Temperature	Confidence Intervals,Wind Speed	Sensors
m	17.96910339	1.290306947	A
m-h	17.81214113	1.246227039	
m+h	18.12606565	1.334386854	
m	18.06542811	1.242124394	B
m-h	17.9047269	1.197166335	
m+h	18.22612932	1.287082454	
m	17.91313662	1.371463217	C
m-h	17.75492624	1.324303789	
m+h	18.07134701	1.418622646	
m	17.99636217	1.581649151	D
m-h	17.83814661	1.529648042	
m+h	18.15457772	1.63365026	
m	18.35393939	0.596242424	E
m-h	18.18193395	0.568059905	
m+h	18.52594484	0.624424943	

Figure 20: 95 percent Cofidence Intervals

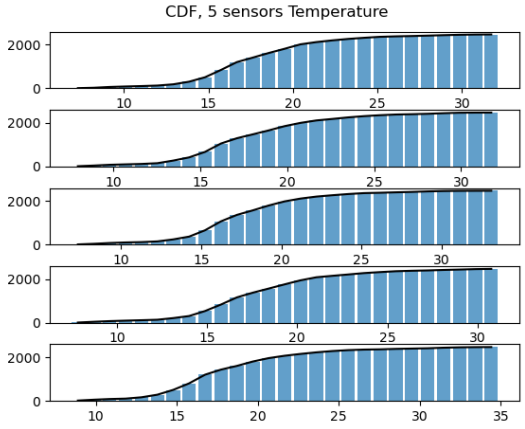

Figure 21: CDF, 5 sensors

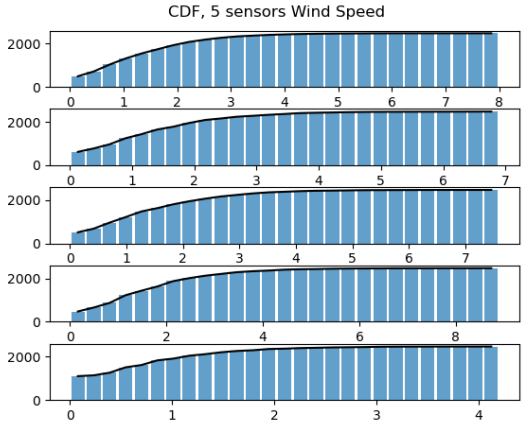

Figure 22: CDF, 5 sensors

4.2 Hypothesis test

Test the hypothesis: the time series for Temperature and Wind Speed are the same for sensors:

1) E, D;
2) D, C;
3) C, B;
4) B, A;

T.test:The null hypothesis over those four cases,returns the t and the p value for each variable(Temperature,Wind Speed).The p value shows which is the probability of seeing the apparent effect if the null hypothesis is true. A p-value higher than 0.05($p>0.05$), is not statistically significant and indicates strong evidence for the null hypothesis. If the p-value is near 0, the effect is said to be statistically significant, which means it is unlikely to have occured by chance. The t.test results are shown in the table underneath.

Variable	Student test	p value	Sensors
Temperature	3.00023	0.00271	ED
Wind Speed	-32.67317	0.00000	ED
Temperature	0.72939	0.46580	DC
Wind Speed	5.87115	0.00000	DC
Temperature	-1.32423	0.18549	CB
Wind Speed	3.89266	0.00010	CB
Temperature	0.84084	0.40048	BA
Wind Speed	-1.50061	0.13352	BA

Figure 23: T.test results

5 Bonus Question

My "employer" wants to estimate the day of maximum and minimum potential energy consumption due to air conditioning usage. To hypothesize regarding those days, I am asked to identify the hottest and coolest day of the measurement time series provided.

To do that, in the beginning, I defined a function that fills a dictionary like this: first column with the mean temperatures of every day data set given, and second column with the days count. The measurement frequency is 20 minutes. Thus, I calculate the mean for every 72 cells of measurements dataset, and I get a final dictionary with 34 mean temperatures, and 34 days,placed proportionally. I name the dictionary to a Dataframe so I can sort it. I sort the dataframe, by the values of the mean temperatures. Thus, in the end I get a sorted dataframe that starts from the hottest to the coolest mean temperature, and the days those mean temperatures took place. I call this function 5 times, for each sensor, and I print the results for the hottest and coolest days. The results are :

```
[25.18333333333333  'june 26']
[14.155555555555557 'june 10']
[24.929166666666667 'june 26']
[14.327777777777776 'june 10']
[24.872222222222224 'june 26']
[14.266666666666667 'june 10']
[24.87500000000000  'june 26']
[14.37083333333333  'june 10']
[25.91111111111111  'june 25']
[14.49027777777778   'july 8']
```

Figure 24: Max-min temperatures

As we observe, the 4 sensors have identical results(june 26',june '10). The sensor E indicates an other result(june '25,july '8), however, there is a really slight difference if we check the sorted dataframe, with the upper results(june 26',june '10). Consequently, I conclude :

hottest day: June 26th
coolest day: June 10th

References

[1] Daniela Maiullari and Clara Garcia Sanchez. Measured Climate Data in Rijsenhout. 8 2020.