Assignment 1

Ioannis Dardavesis (5372666)

1 Introduction

For this exercise, climate data from 5 sensors based in Rijsenhout[1] was used in order to complete its tasks. The sensors have information regarding temperature wind speed and other variables.

2 Lesson A1

2.1 Compute mean statistics (mean, variance and standard deviation for each of the sensors variables), what do you observe from the results?

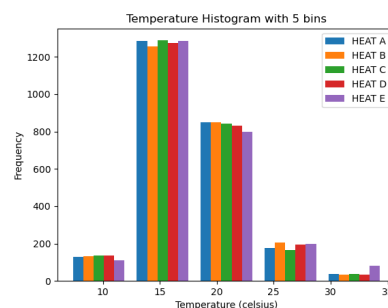| Sensors	Statistics	Direction β True	Wind Speed	Crosswind Speed	Headwind Speed	Temperature	Globe Temperature	Wind Chill	Relative Humidity	Heat Stress Index	Dew Point	Psychro Wet Bulb Temperature	Station Pressure	Barometric Pressure	Altitude	Density Altitude	NA Wet Bulb Temperature	WBGT	TWL	Direction 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	10105.68
	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	766.55	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	1016.69	1016.65	-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,Variance and std of sensors variables

Regarding the statistics results, there are not significant observations that can be made. There is not one sensor that has the highest or lowest results for every variable. In that case, further analysis is needed in order to make secure observations.

2.2 Create 1 plot that contains histograms for the 5 sensors Temperature values. Compare histograms with 5 and 50 bins, why is the number of bins important?
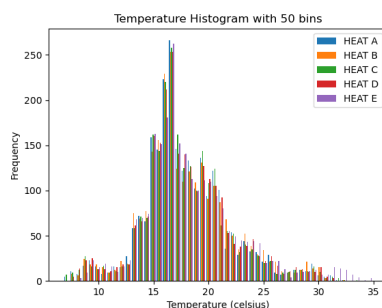
(a) Temperature Histogram (50 bins)   (b) Temperature Histogram (5 bins)

The two figures above show the temperature histograms with 50 and 5 bins respectively. There are some significant differences between the 2 plots. The 50-bins histogram contains

more detail than the 5-bins one. The minimum number of bins, in order to come to robust conclusions about the sample, can be calculated with Rice's rule.

2.3   Create 1 plot where frequency polygons for the 5 sensors Temperature values overlap in different colors with a legend.
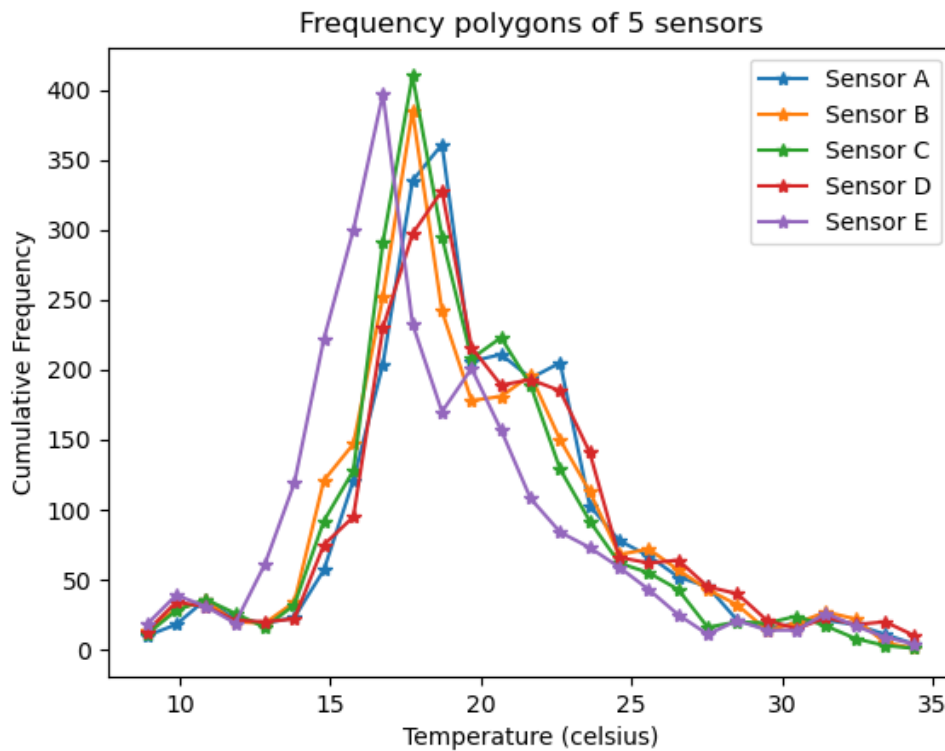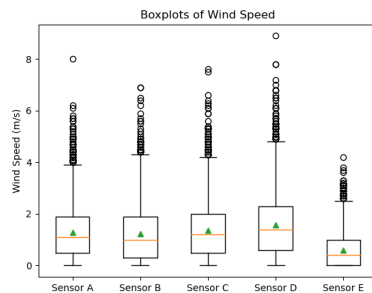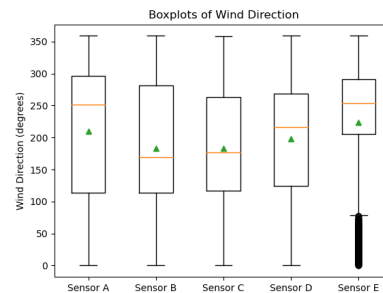

Figure 3: Frequency Polygons of 5 sensors

In order to create the polygons, the number of bins was calculated with Rice's rule. The resulting number was 27.
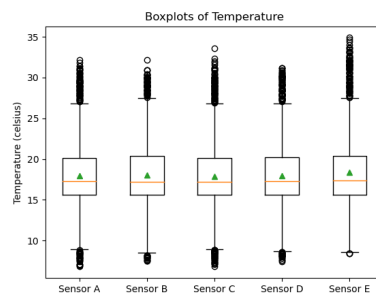
2.4   Generate 3 plots that include the 5 sensors boxplot for: Wind Speed, Wind Direction and Temperature.



(a) Boxplots of Wind Speed

(b) Boxplots of Wind Speed

(c) Boxplots of Wind Speed

Each box represents a sensor.In total, three plots were created one for each variable. In the above figures a lot of outliers can be noticed in the Boxplot of Wind direction, for Sensor E.
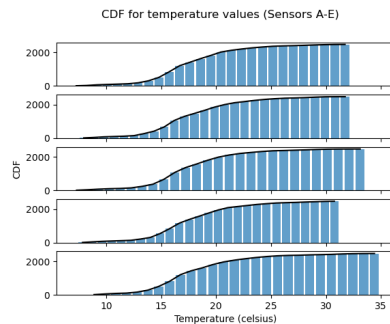
3   Lesson A2

3.1   Plot PMF, PDF and CDF for the 5 sensors Temperature values in independent plots (or subplots). Describe the behavior of the distributions, are they all similar? what about their tails?
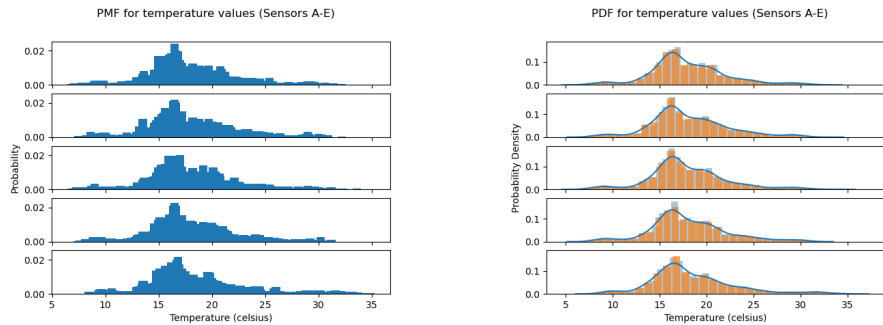

(a) PMF for Temperature values

(b) PDF for Temperature values

(c) CDF for Temperature values

PMF and PPF seem to have similar distributions. Their peaks appear to be at 17-18 celsius, which is the mean temperature for all the sensors.On the other hand, CDF is a cumulative function that maps from a value to its percentile rank. That is why it is noticeable that the graph has higher y values from left to right. In all of the graphs though, temperature values of Sensor E appear to have bigger tails.
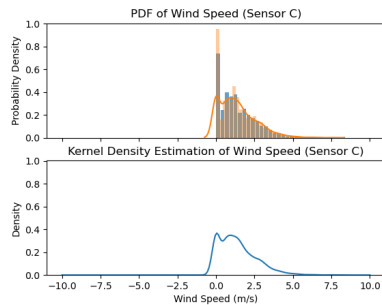
3.2   For the Wind Speed values, plot the pdf and the kernel density estimation. Comment the differences.
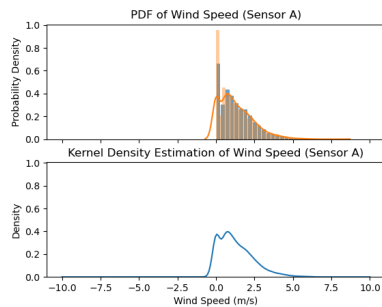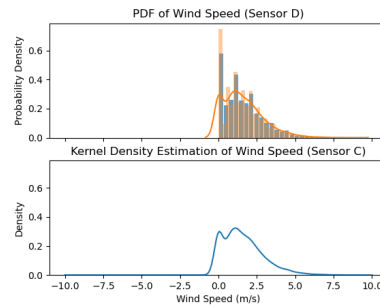

(a) PDF and KDE for Wind Speed values (A)
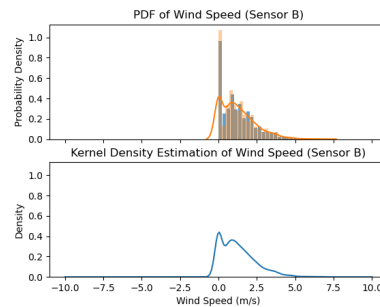
(b) PDF and KDE for Wind Speed values (B)

(c) PDF and KDE for Wind Speed values (C)

(d) PDF and KDE for Wind Speed values (D)
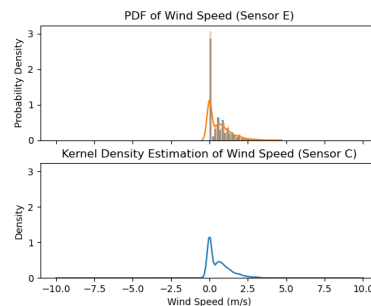
(e) PDF and KDE for Wind Speed values (E)

In order to answer to this question, five figures were made, one for each sensor. There are two plots per figure, one representing the Probability Density Function and the other one the Kernel Density Estimation for the wind speed values. The two graphs are not the same but they have minor differences in each sensor. This result has a logical explanation, because KDE is an algorithm that takes a sample and finds an appropriately smooth PDF that fits the data.

4   Lesson A3

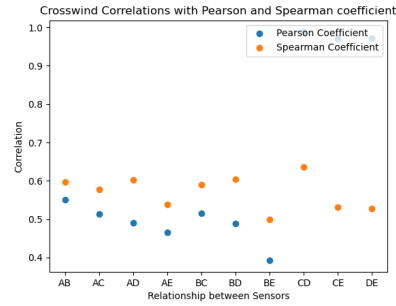## 4.1   Compute the correlations between all the sensors for the variables: Temperature, Wet Bulb Globe Temperature (WBGT), Crosswind Speed. Perform correlation between sensors with the same variable, not between two different variables; for example, correlate Temperature time series between sensor A and B. Use Pearson's and Spearmann's rank coefficients. Make a scatter plot with both coefficients with the 3 variables.

| Sensors Relationship	Variables	Pearson Coefficient	Spearman coefficient
AB	Temperature	0.98810	0.98738
AB	Crosswind	0.55035	0.59698
AB	WBGT	0.99126	0.99213
AC	Temperature	0.98861	0.98829
AC	Crosswind	0.51405	0.57723
AC	WBGT	0.99190	0.99247
AD	Temperature	0.98561	0.98463
AD	Crosswind	0.48990	0.60189
AD	WBGT	0.98701	0.98829
AE	Temperature	0.96920	0.97177
AE	Crosswind	0.46512	0.53784
AE	WBGT	0.94983	0.94913
BC	Temperature	0.98449	0.98544
BC	Crosswind	0.51610	0.59068
BC	WBGT	0.98973	0.98986
BD	Temperature	0.98627	0.98605
BD	Crosswind	0.48803	0.60482
BD	WBGT	0.98786	0.98737
BE	Temperature	0.97209	0.97686
BE	Crosswind	0.39215	0.50028
BE	WBGT	0.95441	0.95690
CD	Temperature	0.98874	0.98819
CD	Crosswind	0.56289	0.63591
CD	WBGT	0.99182	0.99142
CE	Temperature	0.97210	0.97734
CE	Crosswind	0.47323	0.53223
CE	WBGT	0.94927	0.94935
DE	Temperature	0.97137	0.97585
DE	Crosswind	0.46519	0.52733
DE	WBGT	0.94809	0.94870

Figure 7: Pearson and Spearmann Correlation between the sensors
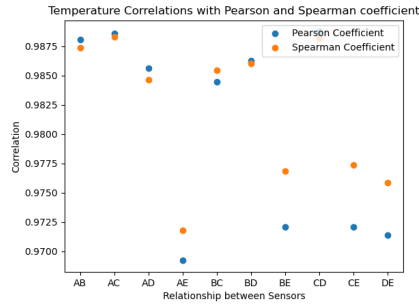
(a) Temperature correlation between (b) Crosswind correlation between sensors
sensors   sors
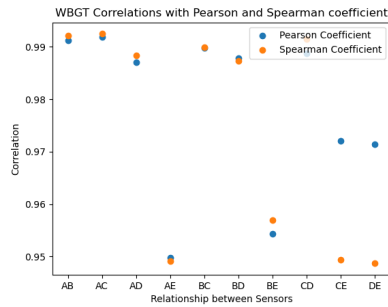
(c) WBGT correlation between sensors

4.2   What can you say about the sensors' correlations?

By comparing the matrix and the scatter plots, some results on the sensors' correlations can be presented. First of all, the correlation values from the Crosswind data are lower, close to 0.5, compared to the ones produced by the other two variables.In the Temperature and WBGT plots all the sensors have very high correlations between each other. There are differences between Pearson and Spearman Coefficients, a fact that is caused due to their different abilities in expressing linear and non-linear information. It is detectable that in all the plots, Sensor E involved in the lowest correlations.

4.3   If we told you that that the sensors are located as follows, hypothesize which location would you assign to each sensor and reason your hypothesis using the correlations.


Figure 9: Sensor Map

Depending on the correlation values between the sensors, as well as the scatter plots, some conclusions were made.After the creation of some scatter plots for the three variables,it was found that the temperature and WBGT correlations were linear, while the scatter plot of croswind was non linear. That is the reason, the Spearman Coefficient was prefered in order to justify the location of the sensors based on crosswind. First of all, in my opinion Sensor E is on the northeast side of the map as shown above. This conclusion was made, due to the fact that E has the lowest correlations with the other sensors. From the location of the sensor, it can be observed that it is surrounded by buildings,in contrast to the other locations, which may explain the low correlation regarding windspeed. Moreover, next to it Sensor C was placed, due to the fact that it has the least worse correlations to Sensor E, comparing to the other sensors. South of C, Sensor D was assigned,because these 2 sensors have the best correlations in terms of WindSpeed and Temperature. Their location is close, they seem to be based on the same type of soil and surrounded by the same buildings. West of D, Sensor B was assigned because comparing to A it is better correlated to D, in terms of all the variables.In the end, Sensor A was placed in the southwestern part, because it is really close to B and also surrounded by the same buildings, a conclusion that can be made due to their high correlations in the three variables.

5　Lesson A4

5.1　Plot the CDF for all the sensors and for variables Temperature and Wind Speed, then compute the 95 confidence intervals for variables Temperature and Wind Speed for all the sensors and save them in a table (txt or csv form).
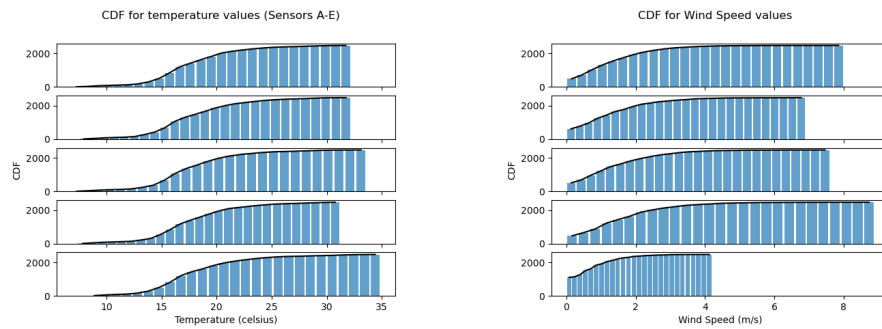
(a) CDF for Temperature values　　　　　(b) CDF for Wind Speed values

It is noticeable that Sensor E has the biggest tail for Temperature values. At the same time, it appears to have the shortest one for Wind speed.

Variables	Confidence Intervals			Sensors
	m-h	m	m+h	
Temperature	17.8121	17.9691	18.1261	A
	17.9047	18.0654	18.2261	B
	17.7549	17.9131	18.0713	C
	17.8381	17.9964	18.1546	D
	18.1819	18.3539	18.5259	E
Wind Speed	1.2462	1.2903	1.3344	A
	1.1972	1.2421	1.2871	B
	1.3243	1.3715	1.4186	C
	1.5296	1.5816	1.6337	D
	0.5681	0.5962	0.6244	E

Figure 11: Confidense intervals between Sensors

5.2 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. What could you conclude from the p-values?

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

Figure 12: Student test and p-values

The hypothesis for the requested sensors was made and the results can be seen above. In order to understand the significance of p-values, a null hypothesis was suggested, that the sensors mentioned above have the same Temperature and Windspeed values. Low p-values indicate that the effect is said to be statistically significant, which means that it is unlikely to have occurred by chance. In that case, the null hypothesis is invalid. The lowest values are observed between Sensors E and D, as well as D and C and they are very close to 0.

6 Bonus question

6.1 Your "employer" wants to estimate the day of maximum and minimum potential energy consumption due to air conditioning usage. To hypothesize regarding those days, you are asked to identify the hottest and coolest day of the measurement time series provided. How would you do that? Reason and program the python routine that would allow you to identify those days.

For the following question the hottest and coolest days per sensor were calculated. In order to achieve that, a function in python was created, entitled as average temperature that took one variable which was the data. The function contains two lists, one named "temperature" and the other one "dates". The first list reads the temperature data from the excel files [1]. Every day there were 72 measurements, every 20 minutes. So, the mean temperature for each day was calculated. The second list includes the strings of the dates, that measurements took place. Furthermore, a dataframe including those two lists was created and then it was sorted from the hottest to the coolest temperature.Then the function prints the hottest and coolest values and the corresponding dates. Last but not least, outside the function, five print statements were coded, so that the results for each sensor can be seen. The results can be seen below:

Sensors	Hot/Cold	Temperature (Celsius)	Date
A	hottest	25.1833	6/26/2020
A	coolest	14.1556	6/10/2020
B	hottest	24.9292	6/26/2020
B	coolest	14.3278	6/10/2020
C	hottest	24.8722	6/26/2020
C	coolest	14.2667	6/10/2020
D	hottest	24.8750	6/26/2020
D	coolest	14.3708	6/10/2020
E	hottest	25.9111	6/25/2020
E	coolest	14.4903	7/8/2020

Figure 13: Highest and lowest temperatures per sensor

For sensors A-D the hottest day is 26/6/2020, while the coolest day 10/6/2020. The difference appears in Sensor E, where the hottest day is 25/6/2020 and the coolest on 8/7/2020.

References

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