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

Homework 1

Özge Tufan (5263719)

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

   Figure 1 illustrates the mean statistics of climate data measured with 5 sensors in different locations within the town of Rijsenhout, during the period 10-06-2020 / 14-07-2020 [Maiullari and Sanchez, 2020]. Although the values are close to each other, the small differences between the sensors are caused by the fact that they are located in different places, and these statistics can help to analyze the reliability of the data to be used in future research and calculations. For this purpose, observing the differences between mean and standard deviation can be the first step to analyze the distribution of the data. For instance, the variables of Crosswind Speed, Headwind Speed, Altitude and Density Altitude are examples where the standard deviation values are higher than that of the mean. This shows that the data for these variables are spread out, which makes the mean value less reliable since there are many outliers affecting the mean. Therefore, it can be concluded that the whole dataset should be analyzed for the above-mentioned variables instead of making a decision based on only the mean values. On the other hand, the remaining variables have smaller standard deviation values than that of the mean for each sensor. For instance, the mean temperature recorded by sensor A is 17.96910339 while the standard deviation is 3.982997522. The fact that the standard deviation is highly smaller than the mean illustrates that the data is clustered around the mean. As a result, the mean temperature can be used to generalize about the weather in the area. Another example can be Relative Humidity, which has a mean of 78.18477383 and a standard deviation of 19.3909788 according to sensor A. This shows that the mean is sufficient to make generalizations instead of having to analyze the whole data.

Variables	Mean Sensor A	Mean Sensor B	Mean Sensor C	Mean Sensor D	Mean Sensor E	Variance Sensor A	Variance Sensor B	Variance Sensor C	Variance Sensor D	Variance Sensor E	Standard Deviation Sensor A	Standard Deviation Sensor B	Standard Deviation Sensor C	Standard Deviation Sensor D	Standard Deviation Sensor E
Direction , True	209.4063005	183.4123586	183.5889248	198.3265966	223.9563636	10108.94031	9977.21777	7703.363096	8133.890057	9308.28508	100.5432261	99.8860239	87.7688048	90.18808157	96.47945418
Wind Speed	1.290306947	1.242124394	1.371463217	1.581649151	0.596242424	1.251154492	1.301501586	1.430920058	1.73981677	0.51122678	1.118550174	1.140833724	1.196210708	1.319021141	0.715001245
Crosswind Speed	0.964943457	0.835621971	0.963298302	1.210509297	0.438505051	0.926592764	0.878585108	1.042574802	1.451502935	0.315941984	0.962596886	0.937328709	1.021065523	1.204783356	0.562087168
Headwind Speed	0.163529887	-0.129806139	-0.262894099	-0.300565885	0.194949495	1.034940101	1.256719316	1.271732179	1.232502712	0.319073108	1.017320058	1.121034931	1.127711035	1.110181387	0.564865566
Temperature	17.96910339	18.06542811	17.91313662	17.99636217	18.35393939	15.86426926	16.62906693	16.10453824	16.10559129	19.04313221	3.982997522	4.077875296	4.013046005	4.013177206	4.363843743
Globe Temperature	21.54458805	21.79943457	21.58738884	21.35929669	21.17616162	68.19135252	66.04931685	67.9413047	61.2022528	63.21550264	8.257805551	8.12707308	8.242651558	7.82318687	7.950817734
Wind Chill	17.83820679	17.94592084	17.77299919	17.83536783	18.2940202	16.26444672	17.03582578	16.54112266	16.55685213	19.13706204	4.032920371	4.127447853	4.067077902	4.069011198	4.374592786
Relative Humidity	78.18477383	77.87831179	77.96285368	77.94203719	76.79305051	376.010059	408.6230082	374.622643	389.8560405	406.4944626	19.3909788	20.21442575	19.35517096	19.74477249	20.16170783
Heat Stress Index	17.89959612	18.0042811	17.82825384	17.9216249	18.28642424	14.99684832	15.43915742	15.35625356	15.11764378	18.47524004	3.872576445	3.929269324	3.918705598	3.88814143	4.298283383
Dew Point	13.55387722	13.53085622	13.45812449	13.50860954	13.55878788	9.72347183	9.636518216	10.08414949	10.07188298	9.422585434	3.118248199	3.104274185	3.17555499	3.173623006	3.069623012
Psychro Wet Bulb Temperature	15.2707189	15.29551696	15.19664511	15.26018593	15.40666667	6.944027119	6.770262723	7.239313447	7.044402877	6.997445432	2.6351522	2.601972852	2.690597229	2.654129401	2.645268499
Station Pressure	1016.168255	1016.657027	1016.689329	1016.728011	1016.166101	38.47126661	36.84193443	37.69149142	34.98778359	38.93991345	6.202520988	6.069755714	6.139339657	5.915047218	6.24018537
Barometric Pressure	1016.128433	1016.616478	1016.6519	1016.688884	1016.127798	38.46795084	36.82886775	37.67562316	34.95232686	38.93517684	6.20225369	6.068679243	6.138047178	5.912049294	6.239805833
Altitude	-25.98707593	-30.05815832	-30.33872272	-30.65319321	-25.96121212	2663.641045	2545.708131	2608.534634	2419.723591	2692.353386	51.61047418	50.45501096	51.07381554	49.19068602	51.88789248
Density Altitude	137.3166397	135.5807754	129.6228779	132.4110752	150.84	26510.04435	26863.31024	26986.60297	26516.12573	29714.9275	162.819054	163.9003058	164.2759963	162.8377282	172.380183
NA Wet Bulb Temperature	15.98154281	15.99680937	15.93423605	15.91564268	15.93688889	10.01210768	9.809254462	10.4802791	9.98743414	9.432183526	3.164191473	3.131972934	3.237325918	3.160290199	3.071186013
WBGT	17.25432149	17.32197092	17.22502021	17.17679871	17.18553535	16.13525808	15.83535547	16.54674535	15.5071849	15.48987153	4.016871679	3.979366214	4.067769088	3.937916314	3.93571741
TWL	301.3929321	299.4516963	301.8997575	305.2545675	284.1153131	814.7665642	790.0692214	766.5335139	616.0098073	1289.913383	28.5441161	28.10817001	27.68634165	24.81954486	35.91536416
Direction , Mag	208.9050889	183.2172859	183.0836702	197.8261924	223.8965657	10105.67705	9975.446909	7704.62017	8135.315513	9268.00789	100.5269966	99.8771591	87.77596579	90.19598391	96.27049335

Figure 1: Mean Statistics

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?*

   It is crucial to choose an adequate number of bins to correctly represent the data. In Figure 2, the overall distribution of the temperature data of 5 sensors is illustrated with 5 bins while Figure 3 shows the exact same data with 50 bins. It can be concluded that the histograms in Figure 3 represent the data more successfully since 5 bins are insufficient to show the little changes in the data which can be clearly observed on the tails. For instance, the frequency of 0 - 10 °C interval is shown with only one bin in Figure 2 which misrepresents the data since Figure 3 illustrates that the above-mentioned interval consists of data that fluctuates. Therefore, it is important to determine the number of bins by taking into account the flactuations that should be illustrated.

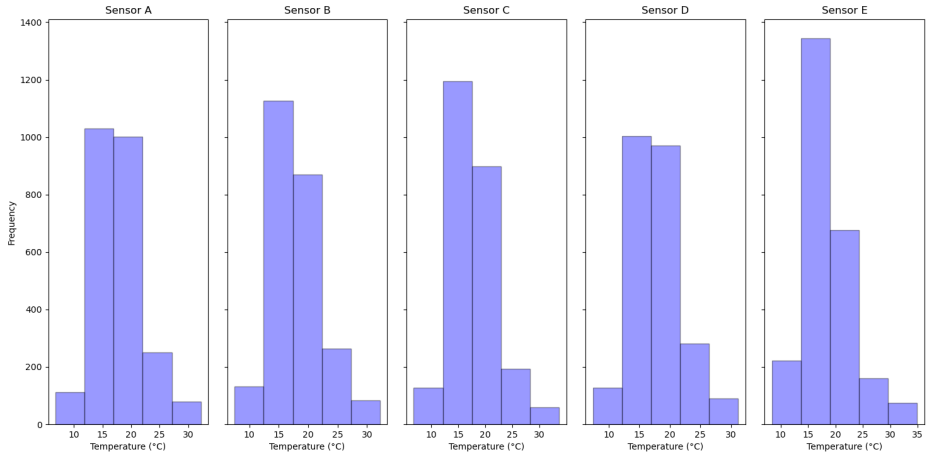

Figure 2: Temperature values with 5 bins

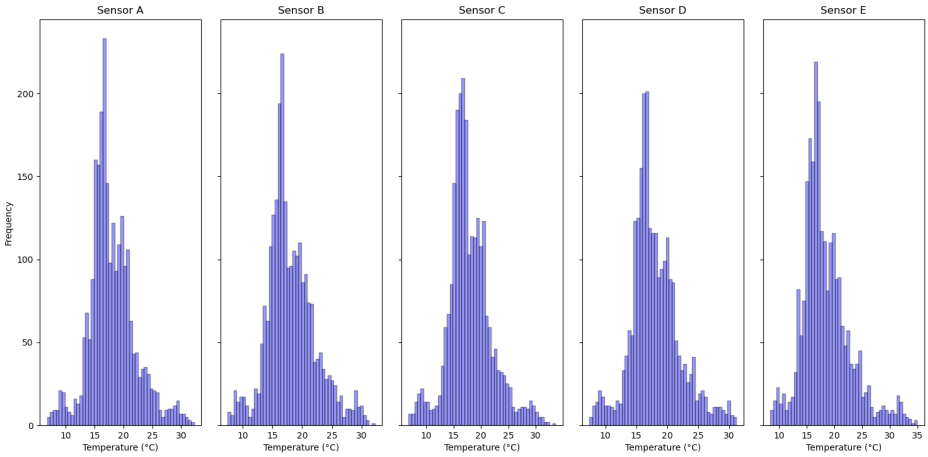

Figure 3: Temperature values with 50 bins

3. *Create 1 plot where frequency poligons for the 5 sensors Temperature values overlap in different colors with a legend.*

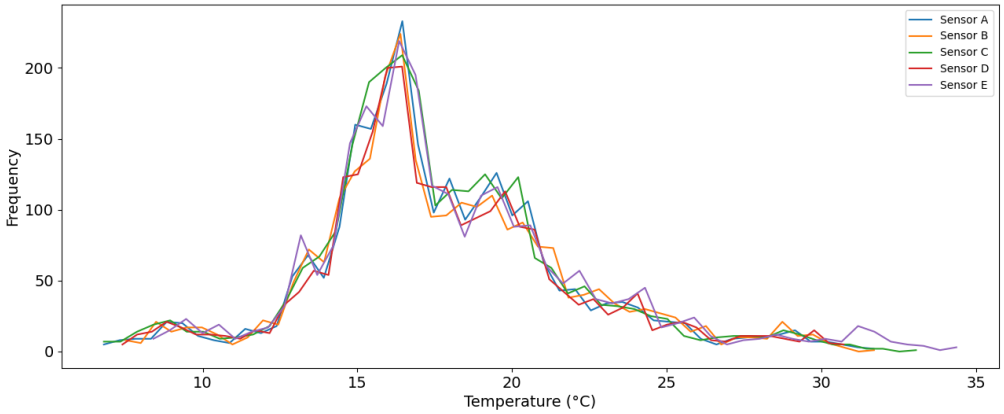

Figure 4: Frequency Polygons

4. *Generate 3 plots that include the 5 sensors boxplot for: Wind Speed, Wind Direction and Temperature.*

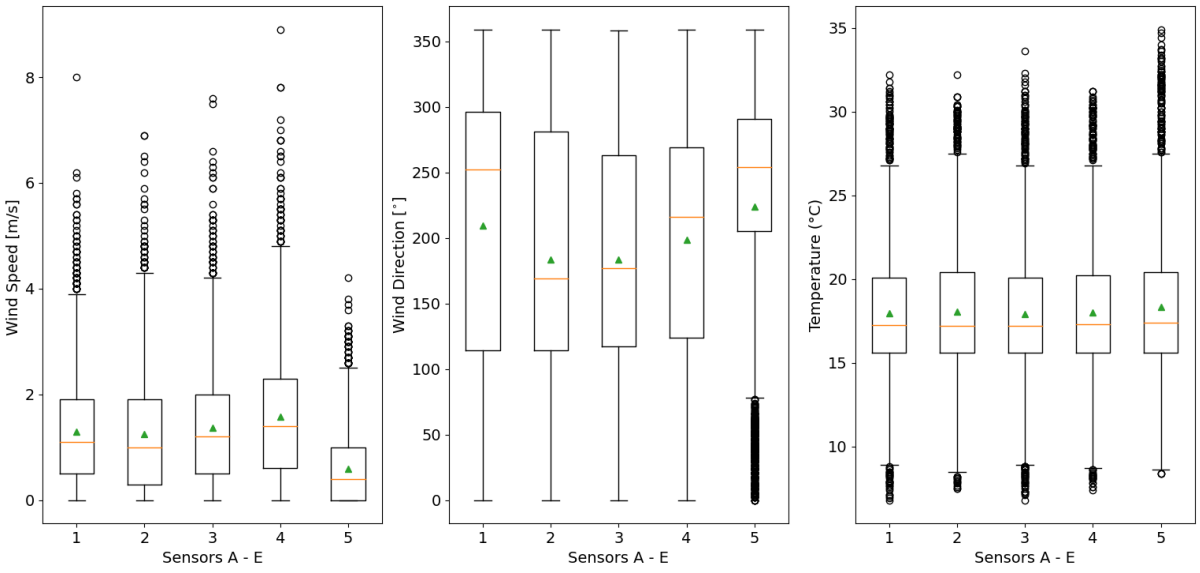

Figure 5: Boxplots for 5 sensors

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

Probability mass functions and probability density functions are beneficial to understand whether the data is distributed evenly or not. In this case, although the graphs are similar, it can be seen that they are right-skewed, which indicates positive skewness. Moreover, cumulative density functions of sensors A to E illustrate that around 50% of the data is smaller than or equal to the mean temperature. This conclusion

can be made by observing the corresponding x-axis value of the y-axis value of 0.5. The mean temperatures recorded by sensors A to E are 17.96910339, 18.06542811, 17.91313662, 17.99636217 and 18.35393939 respectively. Figure 8 shows that the y-axis value of 0.5 corresponds to a value around 17 - 18 °C, which proves the aforementioned assumption.

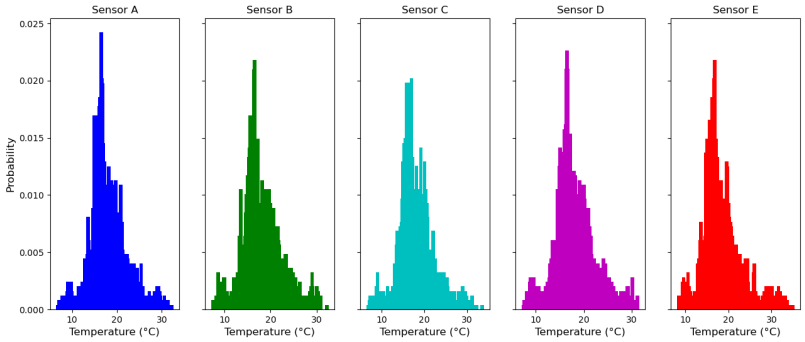

Figure 6: Probability Mass Functions

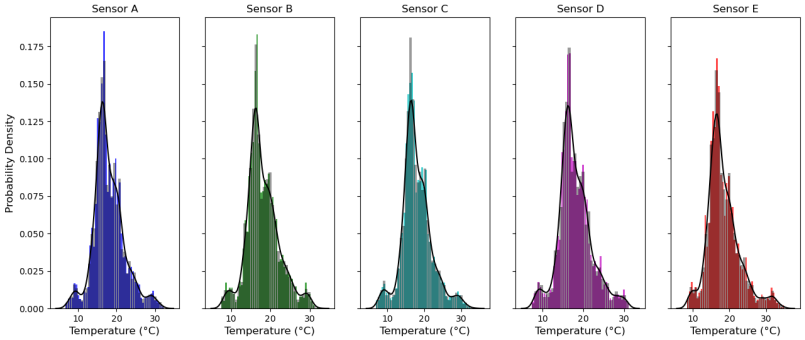

Figure 7: Probability Density Functions

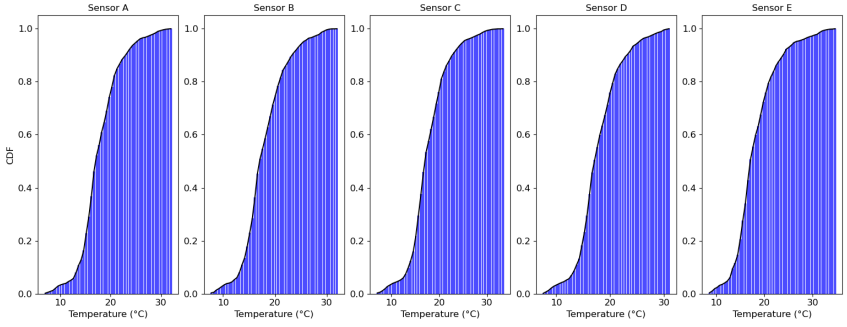

Figure 8: Cumulative Density Functions

6. *For the Wind Speed values, plot the pdf and the kernel density estimation. Comment the differences.*

   Kernel density estimation is used to create a smooth PDF; however, there are still differences between PDF and KDE since KDE takes a sample for this operation which may be insufficient to represent the data at some points. For instance, the peak point of the KDE for Sensor A is around 0.4 while there are higher values shown in the PDF. This is the case for all other sensors as well, which is caused by the fact that KDE makes an estimation of the PDF and this may result in misrepresentations.

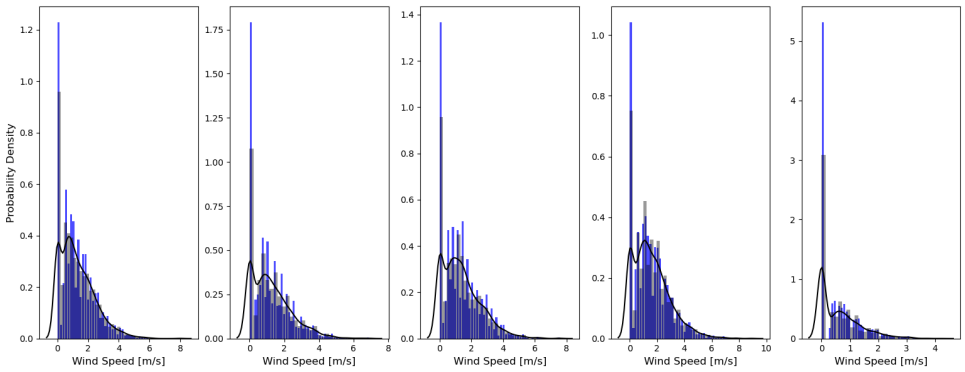

Figure 9: Probability Density Function of Wind Speed

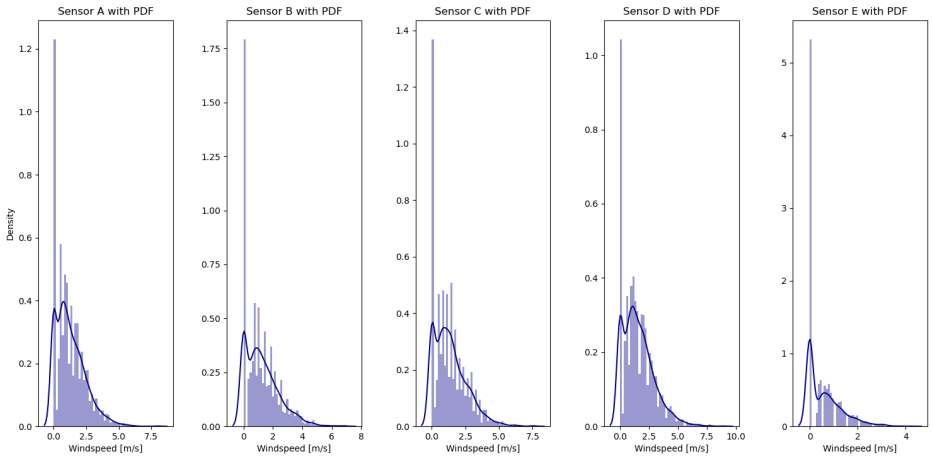

Figure 10: Kernel Density Estimation of Wind Speed

7. *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.*

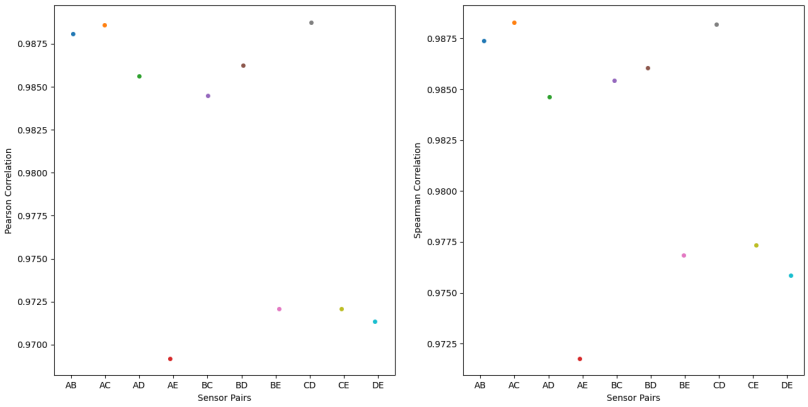

Figure 11: Correlations of Temperature Values

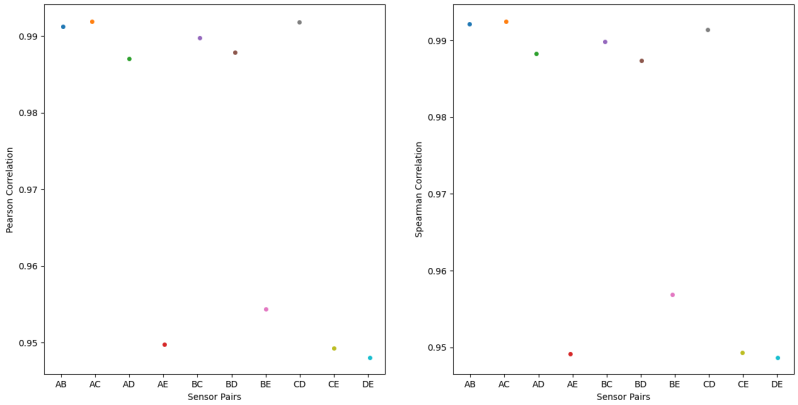

Figure 12: Correlations of Wet Bulb Globe Temperature Values

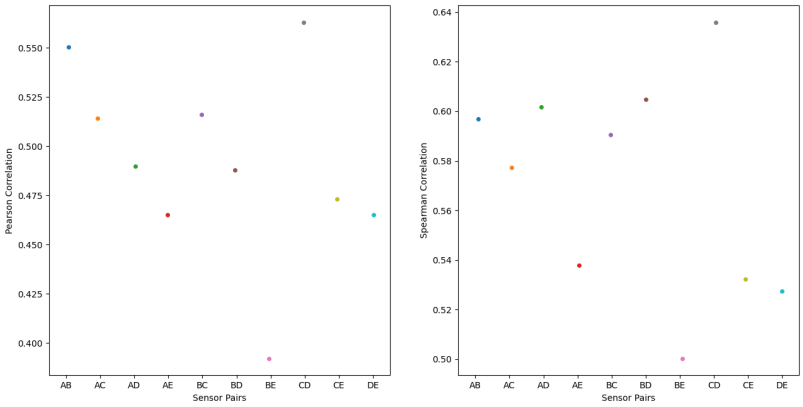

Figure 13: Correlations of Crosswind Speed Values

8. *What can you say about the sensors' correlations?*

   When Pearson's and Spearman's correlation coefficients are calculated for all three variables, it can be seen that the values are relatively low for sensor pairs AE, BE, CE and DE, which means that they have a weak relationship. Since the common sensor in these pairs is E, it can be suggested that the location of sensor E is further away from the other sensors. On the other hand, sensor pairs of AB, AC, AD, BC, BD and CD have a stronger relationship, which can be observed especially in the temperature and wet bulb globe temperature coefficients. The small differences between the coefficients of these sensor pairs may help to determine the location of each sensor. Finally, the coefficients of Crosswind Speed are relatively low for all sensor pairs, the highest being just above 0.550 with sensors C and D. The reason for this may be the fact that Crosswind Speed is affected by many external factors which cannot be degraded to only the locations of the sensors.

9. *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.*

   The correlation coefficients of sensors C and D are the highest in all three variables; therefore, it is more likely that they are located close to each other. Thus, the closest points are assigned to these sensors. When the correlations of the sensor pairs BC and BD are analyzed, it is seen that they both have high values, which concludes to a stronger relationship; therefore, the location on the East of sensor D is assigned to sensor B. Furthermore, it is observed that sensor E has a weak relationship with all other sensors which is interpreted with low coefficients. As a result, the location on the North is assigned to sensor E and the location between sensor B and E is assigned to A.

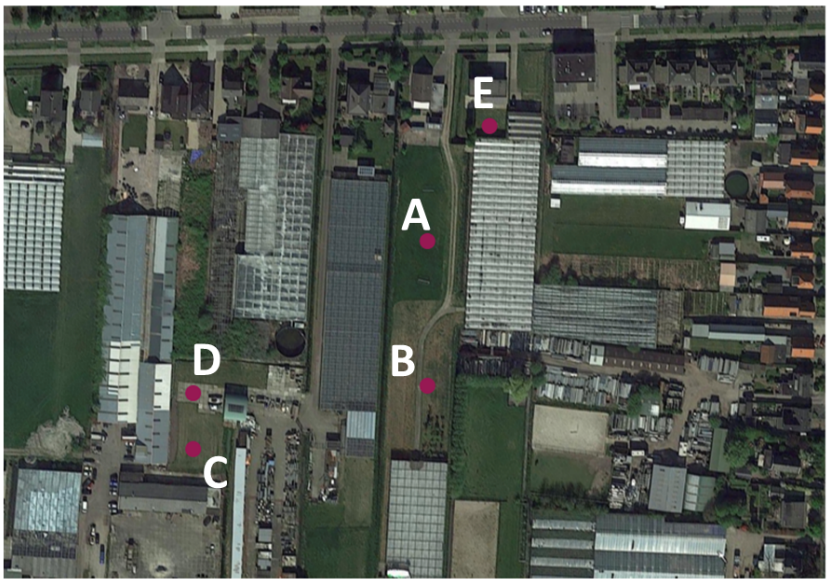

Figure 14: Hypothetical Sensor Locations

10. *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).*

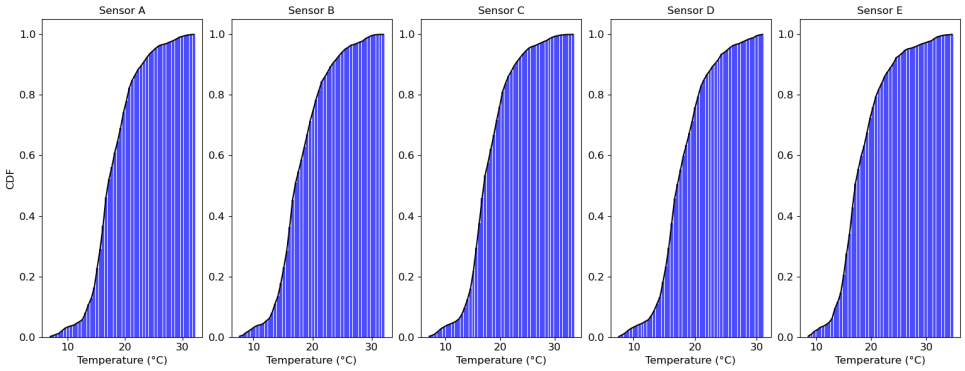

Figure 15: Cumulative Density Functions for Temperature Values

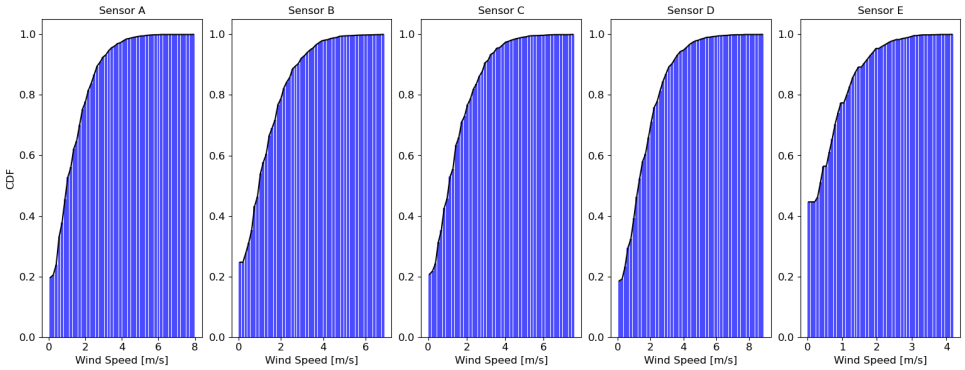

Figure 16: Cumulative Density Functions for Wind Speed Values

Confidence_Intervals - Notepad

File Edit Format View Help

	Sensors	Temperature Confidence Intervals	Wind Speed Confidence Intervals
0	A	(17.81214113267346, 18.126065652463858)	(1.246227038990971, 1.3343868543854427)
1	B	(17.90472689963894, 18.226129320070267)	(1.1971663346979249, 1.287082453670411)
2	C	(17.754926235060246, 18.071347006653575)	(1.3243037885948932, 1.418622646328308)
3	D	(17.83814660824381, 18.15457772482005)	(1.5296480419653757, 1.633650260379006)
4	E	(18.181933946027776, 18.525944841851015)	(0.5680599051948441, 0.6244249432900044)

Figure 17: Confidence Intervals

11. *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.

    Null hypothesis is determined to be a "two-tailed test" because it uses both of the tails of the distribution and is more commonly used in science. Significance level (a) is determined as 0.05 to test the hypothesis and p-values are calculated for each sensor pair.

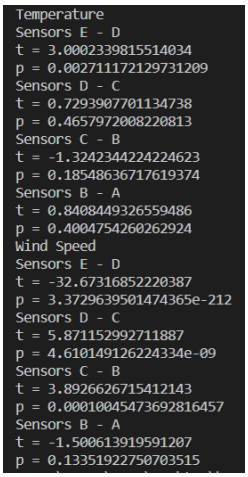

Figure 18: P-values

12. *What could you conclude from the p-values?*

    Firstly, the temperature values of sensor pairs are analyzed for the hypothesis testing. The p-value of sensors E – D is smaller than a; therefore, null hypothesis is rejected. For the other three sensor pairs the p-values are higher than a; as a result, null hypothesis cannot be rejected. Secondly, wind speed values of sensor pairs are analyzed. Here, p-value of sensors B – A is higher than a, which means that null hypothesis cannot be rejected. On the other hand, the p-values of sensors E - D, D - C and C – B are smaller than the significance level, which leads to a rejection of the null hypothesis.

13. *Bonus question: 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 rutine that would allow you to identify those days.*

    The first step is to sort the data according to dates. Each day has 72 recordings; therefore, the mean temperature of every 72 recordings is found. Although the number of recordings differs between sensors, this calculation should be ended in recording 2447 because the last day (07/14/2020) does not have 72 recordings; therefore, it is better to exclude the last day to get a more accurate result. After, a data frame is created with the average temperature values of each day with their corresponding dates. Then, maximum and minimum values are extracted; however, instead of only extracting those values from 'Average Temperature' column, the

rows of the desired values are extracted to be able to see the dates of the hottest and coldest days.

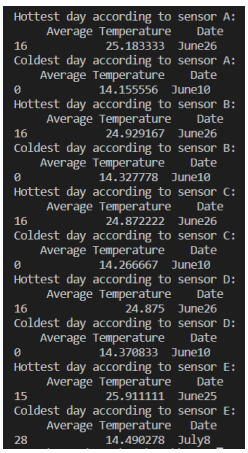

Figure 19: Hottest and coldest days

References

Daniela Maiullari and Clara Garcia Sanchez. Measured Climate Data in Rijsenhout. 8 2020. doi: 10.4121/12833918.v1. URL https://data.4tu.nl/articles/dataset/Measured$_C$$limate_D$$ata_i$$n_R$$ijsenhout$/12833918.