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

Homework 1

Michiel de Jong (4376978)

For the first homework, we have to analyse data[1] from 5 Kestrel sensors. Throughout this report, we will see the results from this analysis. The analysis has been done using Python. A complete repository for the project can be found on `https://github.com/dumigil/geo1001_hw01`

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

	Mean					STD					Var				
	A	B	C	D	E	A	B	C	D	E	A	B	C	D	E
Direction_True	209,406	183,412	183,589	198,327	223,956	100,543	99,886	87,769	90,188	96,479	10108,940	9977,218	7703,363	8133,890	9308,285
Wind_Speed	1,290	1,242	1,371	1,582	0,596	1,119	1,141	1,196	1,319	0,715	1,251	1,302	1,431	1,740	0,511
Crosswind_Speed	0,965	0,836	0,963	1,211	0,439	0,963	0,937	1,021	1,205	0,562	0,927	0,879	1,043	1,452	0,316
Headwind_Speed	0,164	-0,130	-0,263	-0,301	0,195	1,017	1,121	1,128	1,110	0,565	1,035	1,257	1,272	1,233	0,319
Temperature	17,969	18,065	17,913	17,996	18,354	3,983	4,078	4,013	4,013	4,364	15,864	16,629	16,105	16,106	19,043
Globe_Temperature	21,545	21,799	21,587	21,359	21,176	8,258	8,127	8,243	7,823	7,951	68,191	66,049	67,941	61,202	63,216
Wind_Chill	17,838	17,946	17,773	17,835	18,294	4,033	4,127	4,067	4,069	4,375	16,264	17,036	16,541	16,557	19,137
Relative_Humidity	78,185	77,878	77,963	77,942	76,793	19,391	20,214	19,355	19,745	20,162	376,010	408,623	374,623	389,856	406,494
Heat_Stress_Index	17,900	18,004	17,828	17,922	18,286	3,873	3,929	3,919	3,888	4,298	14,997	15,439	15,356	15,118	18,475
Dew_Point	13,554	13,531	13,458	13,509	13,559	3,118	3,104	3,176	3,174	3,070	9,723	9,637	10,084	10,072	9,423
Psychro_WBT	15,271	15,296	15,197	15,260	15,407	2,635	2,602	2,691	2,654	2,645	6,944	6,770	7,239	7,044	6,997
Station_Pressure	1016,168	1016,657	1016,689	1016,728	1016,166	6,203	6,070	6,139	5,915	6,240	38,471	36,842	37,691	34,988	38,940
Barometric_Pressure	1016,128	1016,616	1016,652	1016,689	1016,128	6,202	6,069	6,138	5,912	6,240	38,468	36,829	37,676	34,952	38,935
Altitude	-25,987	-30,058	-30,339	-30,653	-25,961	51,610	50,455	51,074	49,191	51,888	2663,641	2545,708	2608,535	2419,724	2692,353
Density_Altitude	137,317	135,581	129,623	132,411	150,840	162,819	163,900	164,276	162,838	172,380	26510,044	26863,310	26986,603	26516,126	29714,928
NA_WBT	15,982	15,997	15,934	15,916	15,937	3,164	3,132	3,237	3,160	3,071	10,012	9,809	10,480	9,987	9,432
WBGT	17,254	17,322	17,225	17,177	17,186	4,017	3,979	4,068	3,938	3,936	16,135	15,835	16,547	15,507	15,490
TWL	301,393	299,452	301,900	305,255	284,115	28,544	28,108	27,686	24,820	35,915	814,767	790,069	766,534	616,010	1289,913
Direction_Mag	208,905	183,217	183,084	197,826	223,897	100,527	99,877	87,776	90,196	96,270	10105,677	9975,447	7704,620	8135,316	9268,008

Table 1: Mean, Standard Deviation and Variance for all data.

On first glance, the mean values in the table seem to vary quite a bit per sensor, though we cannot say if this is statistically significant. However, it could be an indication that we can gather some good information from these five sensors. The standard deviation values show consistent values across the five sensors. The standard deviation for values related to pressure, which generally does not vary a lot during a day, show a low value consistent with this behaviour, while values related to temperature (which can vary a lot during the day) show much higher standard deviations. The variance, which is the standard deviation squared (actually the other way around), shows values consistent with this mathematical relationship.

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?*

   When comparing the figures below, we can see that although the general shapes of the histograms are rather similar, a lot of information is lost when using only 5 bins. Too many bins however, are also not ideal, as this can cause interference. There are some guidelines to determine the number of bins for a histogram, such as Rice's rule. If we apply Rice's rule to this sample, with Rice's rule being defined as: $k = 2 * \sqrt[3]{n}$.

With n being approximately 2500 for all sensors, we thus get $2 * \sqrt[3]{2500} = 27.144$, which would give us 27 bins. This is the number of bins used for histograms and histogram derived plots in this assignment.

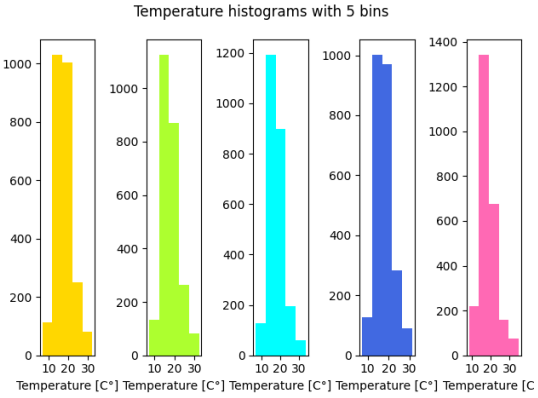

Figure 1: Histogram of five sensors using 5 bins

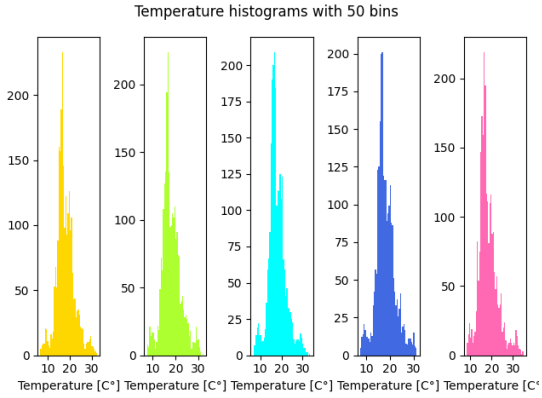

Figure 2: Histogram of five sensors using 50 bins

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

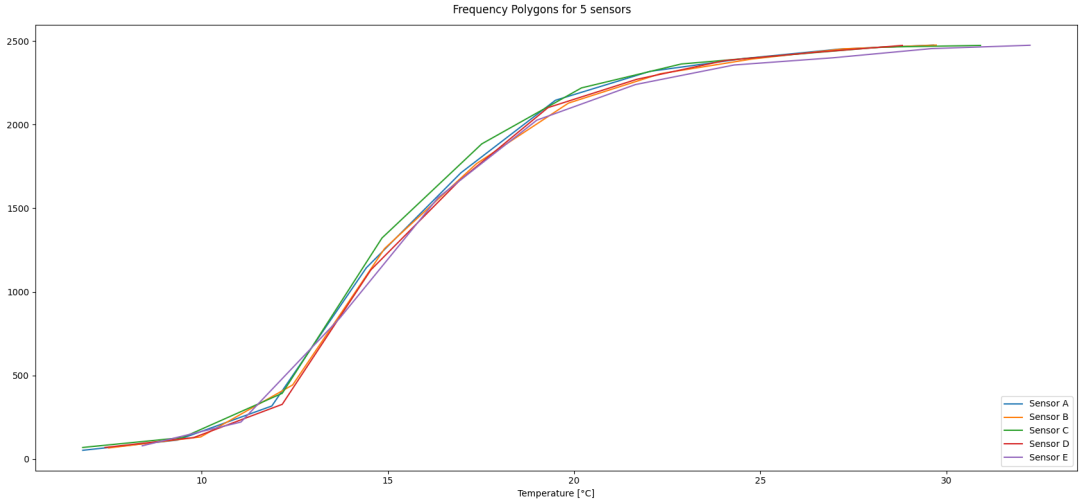

Figure 3: Frequency polygons for the 5 sensors Temperature values

In the figure above, we see that the five sensors have similar frequency polygons. From this we could infer, that they have similar distributions, which would support the observation we already made by looking at the histograms in the previous question.

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

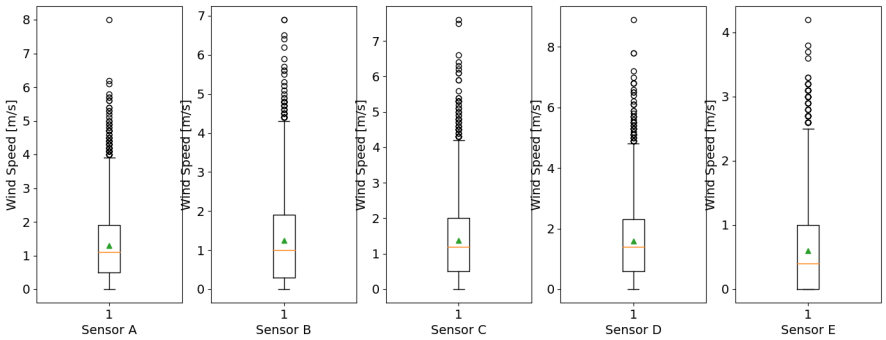

Figure 4: Box plots for the 5 sensors Wind Speed values

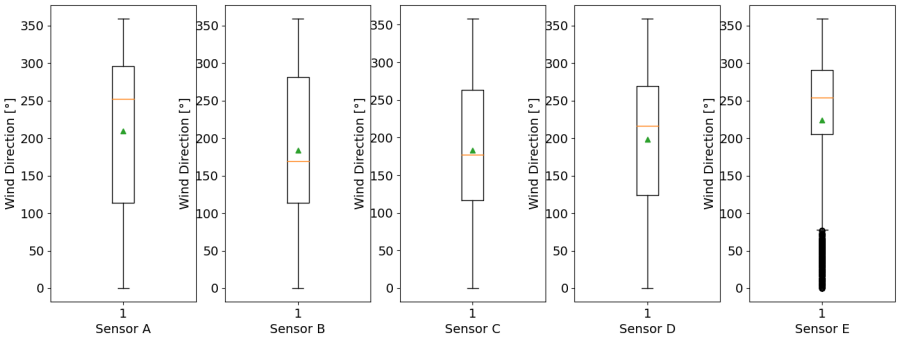

Figure 5: Box plots for the 5 sensors Wind Direction values

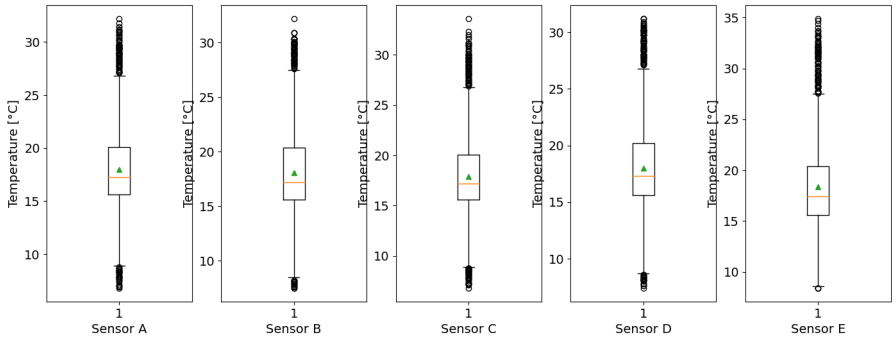

Figure 6: Box plots for the 5 sensors Temperature values

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?*

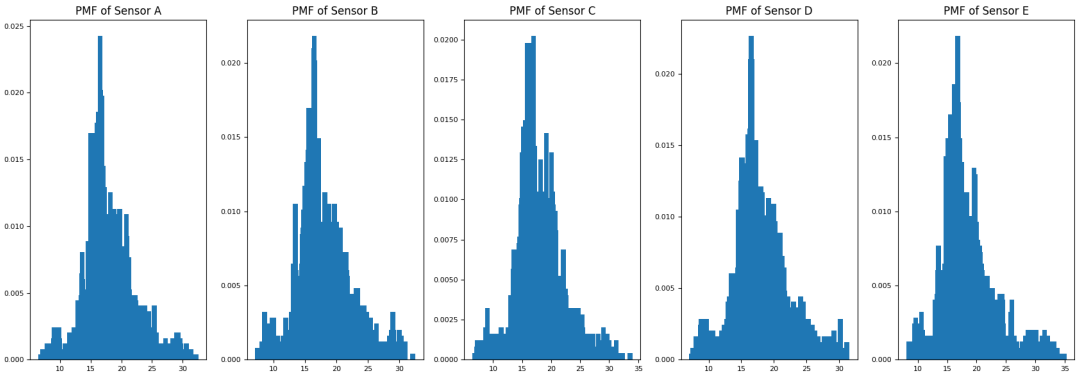

Figure 7: Probability mass function for the 5 sensors Temperature values

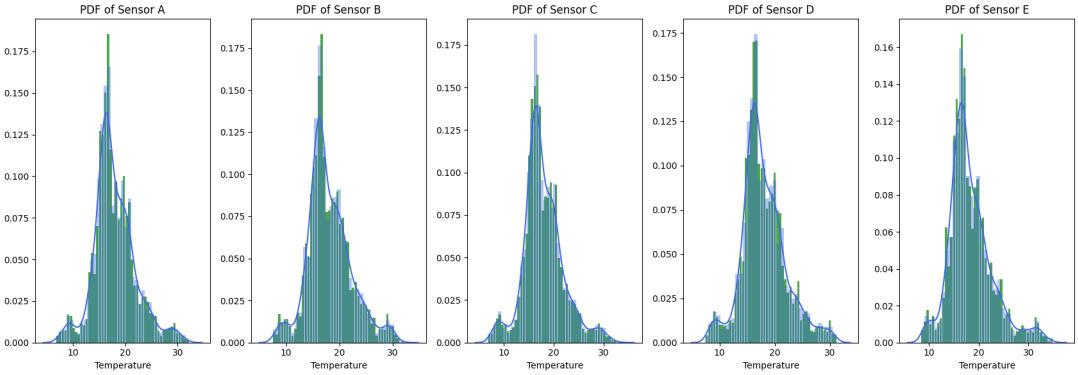

Figure 8: Probability density function for the 5 sensors Temperature values

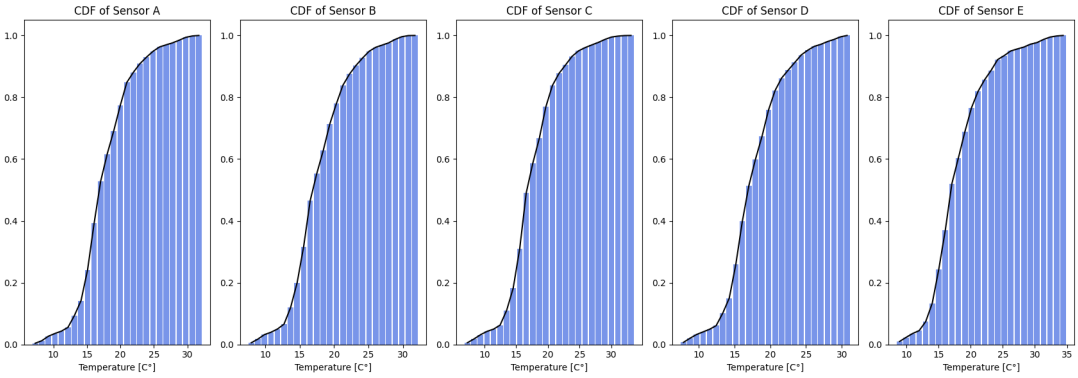

Figure 9: Cumulative density function for the 5 sensors Temperature values

Upon first glance, the PMF and PDF plots look very similar for the five sensors: all five they seem very narrow bell curves, meaning that a lot of the values are centered around the mean. In terms of skewness, all the sensors seem to be skewed slightly to the left, with their peaks around 17-18 degrees celsius. The also have a second mini

peak around 20 degrees celsius. This peak is not present left of the mean, which is an interesting observation. All five sensors have a distinct "mini-peak" in both their tails. In terms of kurtosis, the data seems to present narrow peaks, and fatter tails; indicating high kurtosis. When we look back at table 1, we see that all five sensors have similar standard deviation and variance, indicating similar variability. When looking at the CDF plots, we see our hunch supported, with seeing a higher mass under 0.5 than above it.

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

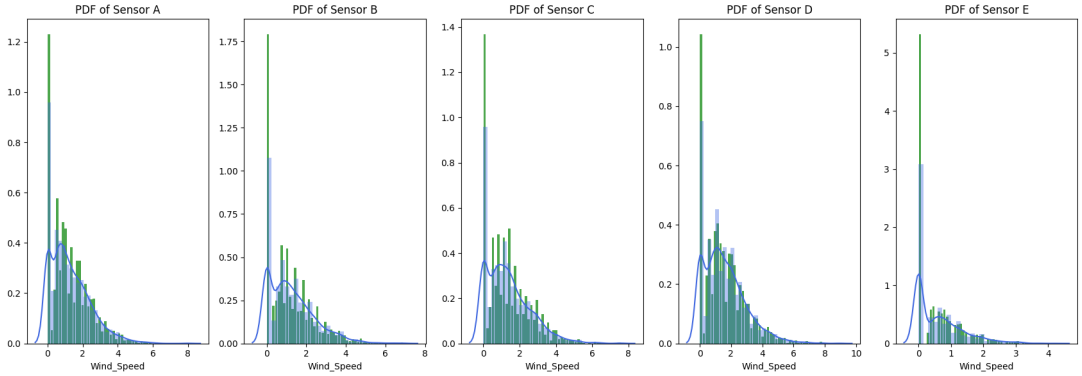

Figure 10: Probability density functions for the sensors' Wind Speed measurements.

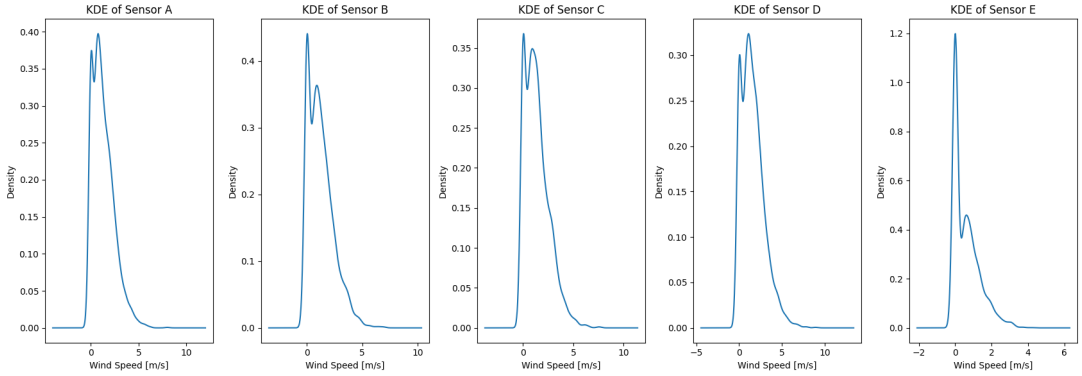

Figure 11: Kernel density estimates for the sensors' Wind Speed measurements.

The PDF's and the KDE's for Wind Speed differ quite a bit in shape between them. When looking at the PDF's we see that the sensors skew to the right, and have a peak close to zero. Sensors A to D are more similar, with E being the odd one out with a very narrow peak, but also a small tail. This behaviour is translated by the KDE's, which seems to broaden the peak for sensors A to D. In those PDF's we see that there is a second peak around 2 m/s. In sensors A to D, the KDE has tried to combine this peak with the main peak. In sensor E, the proportions observed in the PDF are somewhat maintained.

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 Spearman's rank coefficients. Make a scatter plot with both coefficients with the 3 variables.*

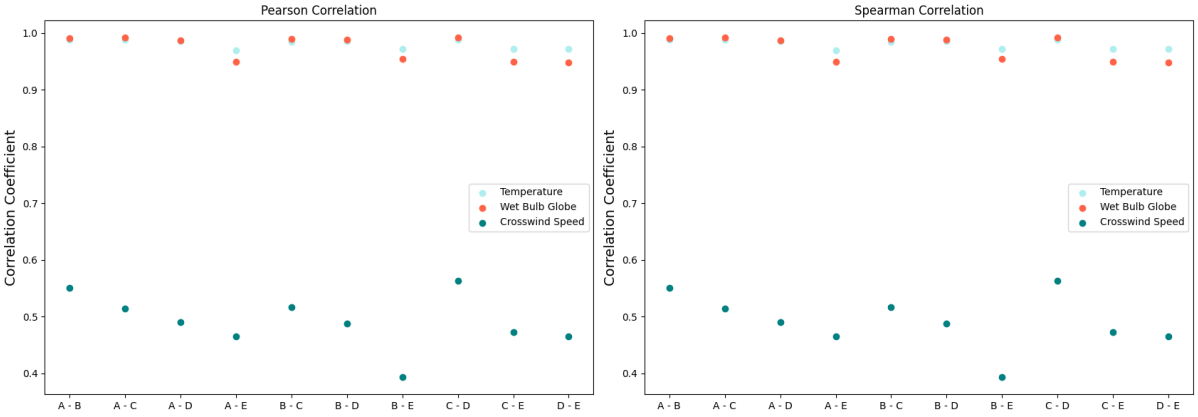

Figure 12: Scatterplot of Pearson and Spearman correlation coefficients for all combinations possible between sensors.

8. *What can you say about the sensors' correlations?*
When looking at the correlations, we can see that in general, the correlations coefficients for the variables "Temperature" and "Wet Bulb Globe Temperature" are really high, indicating a high level of correlation between the sensors. This makes sense, because they both measure temperature, albeit with slightly different parameters. The correlation coefficients for "Crosswind Speed" vary between 0.4 and 0.6, indicating a lower correlation. However, there is much more variability between the sensors, with stronger correlations in some pairs, and lower correlations in others.

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.*

Figure 13: Locations of sensors without labels on a map.

We can see that that lowest correlation values for "Temperature" and "WBGT" are A-E, B-E, C-E and D-E, this would suggest that the temperatures at sensor E vary the most from all the other sensors. Thus, I would hypothesize that sensor E will be the sensor in the top right. When looking at the "Wind Speed" correlations, we see that sensor E is also the odd one out. This would further my hypothesis that E is the sensor in the top right, because it looks to be semi enclosed by walls on three sides.
Furthermore, we can see that the correlations for "Cross Wind" between the pairs A-B and C-D are slightly higher than other pairs. This could indicate that A-B are a pair and C-D are a pair, with sensors C and D being in the center open field, and A and B being on the bottom left. C and D both have higher average "Wind Speed" and "Cross Wind Speed" values, which would be consistent with a large open field. By process of elimination, A and B would thus be located bottom left.

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).*

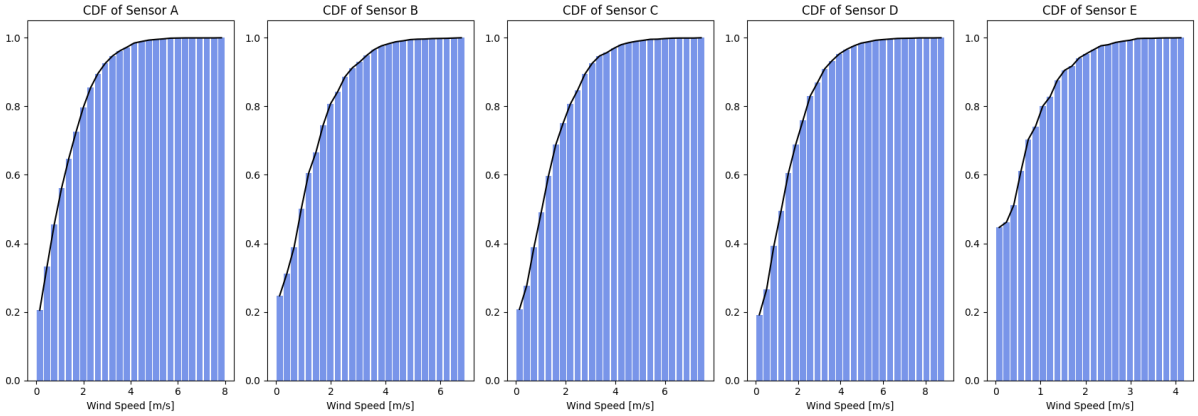

Figure 14: CDF for all the sensors Wind Speed values.

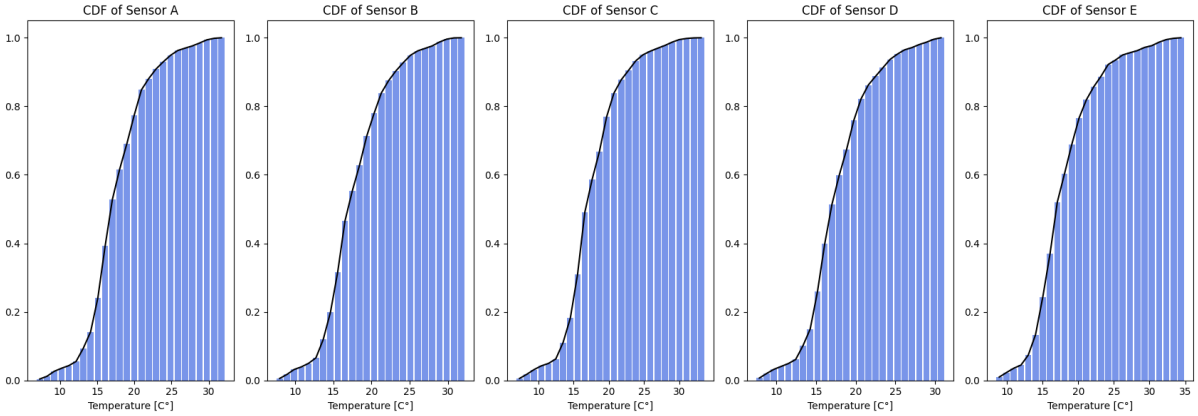

Figure 15: CDF for all the sensors Temperature values.

	Wind Speed			Temperature	
	Lower	Upper		Lower	Upper
A	0	3,484	A	10,156	25,783
B	0	3,479	B	10,066	26,065
C	0	3,717	C	10,044	25,782
D	0	4,168	D	10,127	25,866
E	0	1,999	E	9,795	26,913

Table 2: 95% Confidence intervals for Wind Speed and Temperature value for all five sensors

When computing the 95% confidence intervals for the "Wind Speed" variable, the python script kept computing negative values for the lower bounds. When looking at the CDF for "Wind Speed", this makes sense: there are a lot of values clustered around 0, but of course negative wind speed is not possible. Thus, the lower bounds

have been manually corrected to zero. This has been done in favor of tweaking too much with the interval, only when setting the confidence interval to 40% or lower, did the script compute positive values for the lower bounds.

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

    - *E-D*
    - *D-C*
    - *C-B*
    - *B-A*

	E-D		D-C		C-B		B-A	
	Temperature	Wind Speed	Temperature	Wind Speed	Temperature	Wind Speed	Temperature	Wind Speed
t-value	-3,00023	32,67317	0,72939	5,87115	-1,32423	3,89266	0,84084	-1,50061
p-value	0,00271	0,00000	0,46580	0,00000	0,18549	0,00010	0,40048	0,13352

Table 3: Results of Student's T-test for four pairs of sensors concerning the variables Temperature and Wind Speed.

12. *What could you conclude from the p-values?*
    When looking at the hypothesis for the aforementioned pairs:

$$H_0 : \mu_{sensor1} - \mu_{sensor2} = 0$$

    with a significance level of $\alpha = 0.05$, we can begin to look at the results in table 3.

    For the pair E-D, we can see from the p-values, we can see that both Temperature and Wind Speed differ significantly for sensor E and sensor E. This is not surprising, given our reasoning about the position of sensor E in question 9.
    For the pair D-C, we see that Temperature does not differ significantly, but Wind Speed does. This is an interesting result, because it suggests that one of the sensors is exposed to higher wind speeds than the other. This could support a reasoning for question 9, that the sensor with the higher wind speed values (sensor D) will be the bottom center sensor, because it is not as sheltered as the sensor above it, but it is close, resulting in very similar temperature readings.
    For the pair C-B, we see again that Temperature does not differ significantly, but Wind Speed does. This supports our reasoning in question 9, which does not put the two sensors very close to each other.
    For the pair A-B, we see that both Temperature and Wind Speed do not vary significantly from each other, supporting our reasoning in question 9 which places them in a pair on the bottom left of the map. Because those sensors appear to be closest together of all sensor pairs, this result is not surprising.

13. *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.*

    To identify the hottest and coldest days of the time series, we first have to categorize the time series per day, because there are multiple measurements per day, and then, per day, calculate the mean temperature of that day.

    In python this is done first, by converting the first column string values to a pandas DateTime format, so python will understand we are dealing with dates and times as such:

    ```
    Temperature['DATES'] = pd.to_datetime(Temperature.DATES,
        infer_datetime_format=True)
    ```

    After this, we can try to calculate the mean per day, so we can select the minimum and maximum. We will do this by resampling the data per day, and then calculating the mean per sample as such:

    ```
    Temperature = Temperature.resample('D', on='DATES').mean()
    ```

    This has given us the means of every day in the time series per sensor. Now we want to know the overall mean of the day. We will calculate the means per day of all sensors as such:

    ```
    Temperature = Temperature.mean(axis=1)
    ```

    This has given us the means per day. Now, it is a matter of finding the minimum and maximum temperature and printing them with their associated days.

    ```
    coldest_day = Temperature.idxmin()
    hottest_day = Temperature.idxmax()
    print('The coldest day in the dataset is: ' + str(coldest_day))
    print('The hottest day in the dataset is: ' + str(hottest_day))
    ```

    Which prints the following to the terminal:

    ```
    The coldest day in the dataset is: 2020-06-10 00:00:00
    The hottest day in the dataset is: 2020-06-26 00:00:00
    ```

    From this we can conclude that the coldest day in the time series is the 10^{th} of June 2020 and the hottest day in the time series is the 26^{th} of June 2020.

References

[1] Daniela Maiullari and Clara Garcia Sanchez. "Measured Climate Data in Rijsenhout". In: (Aug. 2020). DOI: 10.4121/12833918.v1. URL: https://data.4tu.nl/articles/dataset/Measured_Climate_Data_in_Rijsenhout/12833918.