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

Homework 1

Louise Spekking (4256778)

The data used in this analysis is data from five Kestrel 5400 sensors placed in the town of Rijsenhout, The Netherlands over the a period from June 10th to July 17th [1].

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

   Firstly, the mean, the variance and the standard deviation of all 19 numeric variables were calculated, for report clarity not all individual variables will be discussed.

Table 1: Means calculated for each measurement per sensor.

Variable	*Sensor A*	*Sensor B*	*Sensor C*	*Sensor D*	*Sensor E*
Wind Direction, True	209.406	183.412	183.589	198.337	223.965
Wind Speed	1.290	1.242	1.371	1.582	0.596
Crosswind Speed	0.965	0.836	0.963	1.211	0.439
Headwind Speed	0.164	-0.130	-0.263	-.0301	0.195
Temperature	17.969	18.065	17.913	17.996	18.354
Globe Temperature	21.545	21.799	21.587	21.359	21.176
Wind Chill	17.838	17.946	17.773	17.835	18.294
Relative Humidity	78.185	77.878	77.963	77.942	76.793
Heat Stress Index	17.900	18.004	17.828	17.921	18.386
Dew Point	13.553	13.531	13.458	13.509	13.559
Psychro Wet Bulb Temperature	15.271	15.296	15.197	15.260	15.407
Station Pressure	1016.168	1016.657	1016.689	1016.728	1016.166
Barometric Pressure	1016.128	1016.616	1016.651	1016.689	1016.128
Altitude	-25.987	-30.058	-30.338	-30.653	-25.961
Station Pressure	137.317	135.581	129.623	132.411	150.840
NA Wet Bulb Temperature	15.981	15.997	15.934	15.916	15.937
WBGT	17.254	17.322	17.225	17.177	17.186
TWL	301.393	299.452	301.900	305.255	284.115
Wind Direction, Mag	208.905	183.217	183.084	198.826	223.987

Table 2: Standard deviations calculated for each measurement per sensor.

Variable	*Sensor A*	*Sensor B*	*Sensor C*	*Sensor D*	*Sensor E*
Wind Direction, True	100.523	99.866	87.751	90.17	96.46
Wind Speed	1.118	1.141	1.196	1.319	0.715
Crosswind Speed	0.962	0.937	1.021	1.205	0.562
Headwind Speed	1.017	1.121	1.127	1.11	0.565
Temperature	3.982	4.077	4.012	4.012	4.363
Globe Temperature	8.256	8.125	8.241	7.822	7.949
Wind Chill	4.032	4.127	4.066	4.068	4.374
Relative Humidity	19.387	20.21	19.351	19.741	20.158
Heat Stress Index	3.872	3.928	3.918	3.887	4.297
Dew Point	3.118	3.104	3.175	3.173	3.069
Psychro Wet Bulb Temperature	2.635	2.601	2.69	2.654	2.645
Station Pressure	6.201	6.069	6.138	5.914	6.239
Barometric Pressure	6.201	6.067	6.137	5.911	6.239
Altitude	51.6	50.445	51.063	49.181	51.877
Density Altitude	162.786	163.867	164.243	162.805	172.345
NA Wet Bulb Temperature	3.164	3.131	3.237	3.16	3.071
WBGT	4.016	3.979	4.067	3.937	3.935
TWL	28.538	28.102	27.681	24.815	35.908
Wind Direction, Mag	100.507	99.857	87.758	90.178	96.251

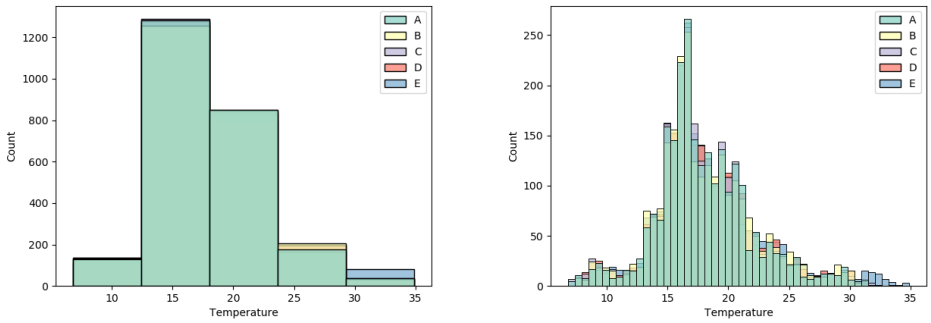
Table 3: Variance deviations calculated for each measurement per sensor.

Variable	*Sensor A*	*Sensor B*	*Sensor C*	*Sensor D*	*Sensor E*
Wind Direction, True	10108.94	9977.218	7703.363	8133.89	9308.285
Wind Speed	1.251	1.302	1.431	1.74	0.511
Crosswind Speed	0.927	0.879	1.043	1.452	0.316
Headwind Speed	1.035	1.257	1.272	1.233	0.319
Temperature	15.864	16.629	16.105	16.106	19.043
Globe Temperature	68.191	66.049	67.941	61.202	63.216
Wind Chill	16.264	17.036	16.541	16.557	19.137
Relative Humidity	376.01	408.623	374.623	389.856	406.494
Heat Stress Index	14.997	15.439	15.356	15.118	18.475
Dew Point	9.723	9.637	10.084	10.072	9.423
Psychro Wet Bulb Temperature	6.944	6.77	7.239	7.044	6.997
Station Pressure	38.471	36.842	37.691	34.988	38.94
Barometric Pressure	38.468	36.829	37.676	34.952	38.935
Altitude	2663.641	2545.708	2608.535	2419.724	2692.353
Density Altitude	26510.044	26863.31	26986.603	26516.126	29714.928
NA Wet Bulb Temperature	10.012	9.809	10.48	9.987	9.432
WBGT	16.135	15.835	16.547	15.507	15.49
TWL	814.767	790.069	766.534	616.01	1289.913
Wind Direction, Mag	10105.677	9975.447	7704.62	8135.316	9268.008

Please type your answer here.

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?*

   In figure 1 two histograms of the temperature measurements are shown. When comparing the two histograms the importance of the number of bins can be observed. Details about the distribution of the data points is better visible in the histogram with 50 bins, whereas in the histogram with 5 bins many nuances of the distribution are lost. However, choosing too many bins will lead to noise and make interpeting the distribution of the data more complicatex. To estimate the needed number of bins Rice's rule of thumb can be used, $\sqrt[3]{N}*2$, in this dataset for temperature values Rice's rule will give $\sqrt[3]{2746} * 2 \approx 27$ bins, meaning that whereas 5 bins is too little to effectively display the data distribution and that 50 is probably too many.



(a) Histogram of temperature (°C) per sensor with 5 bins.

(b) Histogram of temperature (°C) per sensor with 50 bins.

Figure 1: Histograms of the temperature measured by all 5 sensors in °C, with 5 and 50 bins.

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

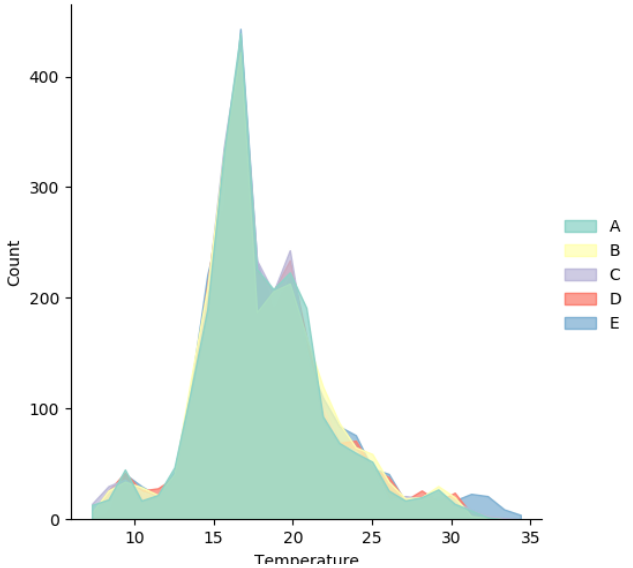


Figure 2: Frequency polygon with 27 bins of the temperature (°C) measured by all 5 sensors.

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

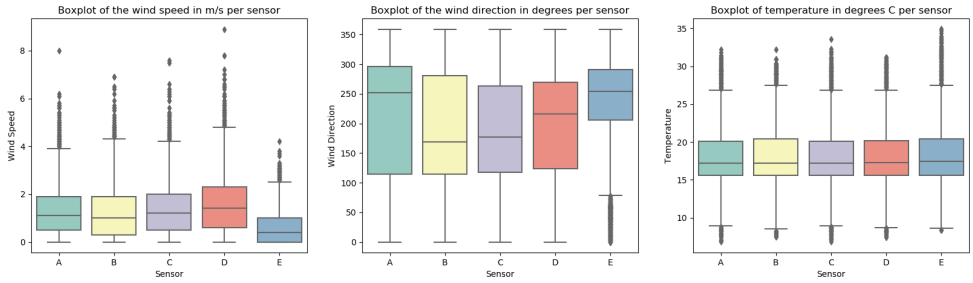


Figure 3: Boxplots showing the data distributions for wind speed (m/s), wind direction (deg) and temperature (°C).

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?*

   The distribution plotted in figures 4, 5, 6 show the distributions of temperature data per sensor. In the first graphs, figure 4, showing the PMF, it can be observed that all 5 sensors are slightly right skewed, and have long tails to the left. In these plots it is also observable that sensor E has the most outliers to the left of the distribution. The PDF plots, figure 5, are very similar to the PMF plots, but outliers in the data have less effect, making it more apparent that the distributions are slightly right skewed with tails to the left. This tail is, similar to the observations in the PMF plot, longest for sensor E. In the bottom plot, figure 6, the CDF plot, the curve resembles an S-curve indicating a relatively normal distribution, however the first part of the curve is relatively flat, indicating only few low temperature values, and the line had a minor kink, indicating a change in frequency of the datapoints.

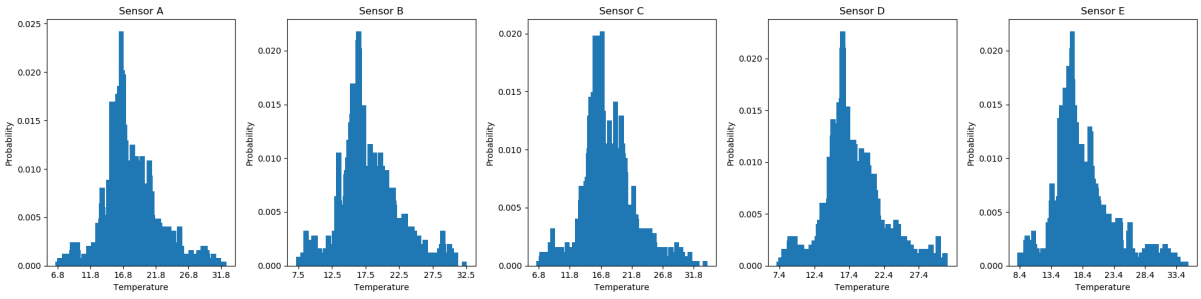


Figure 4: Probability Mass Function of the temperature (°C) values per sensor.

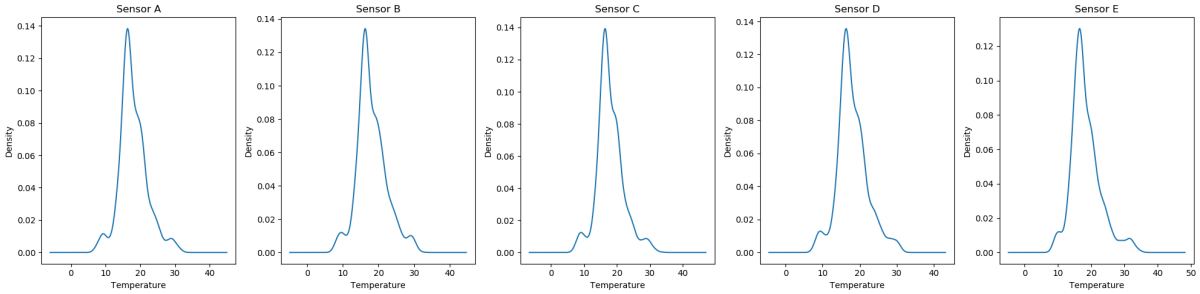


Figure 5: Probability Density Function of the temperature values per sensor.

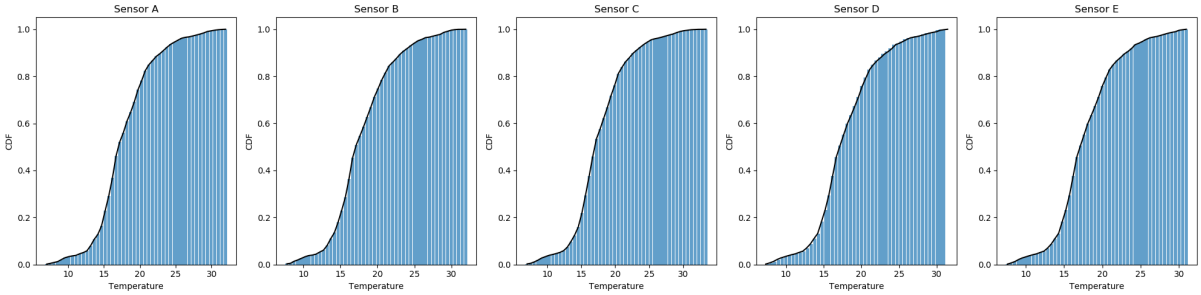


Figure 6: Cumulative Density Function of the temperature (°C) values per sensor.

6. *For the Wind Speed values, plot the pdf and the Kernel Density Estimation. Comment the differences*

   TEXT

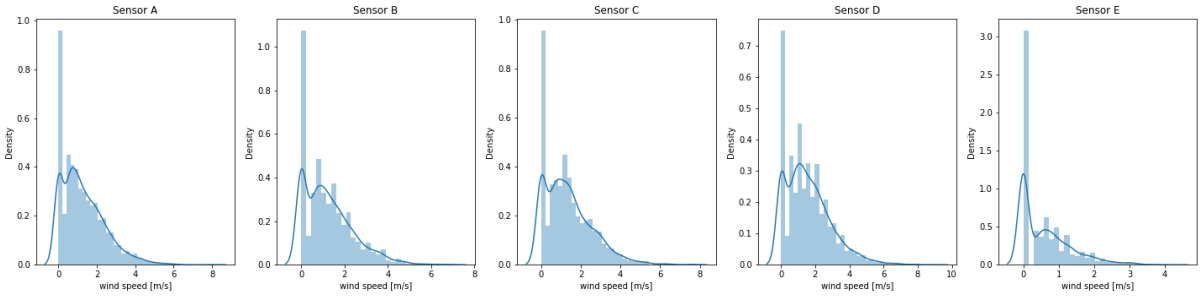


Figure 7: Probability Density Function (bars) plotted with the kernel density estimation (line) of the wind speed (m/s) measurements per sensor

7. *Compute the correlations between all the sensors for the variables: Temperature, Wet Bulb Globe, 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.*

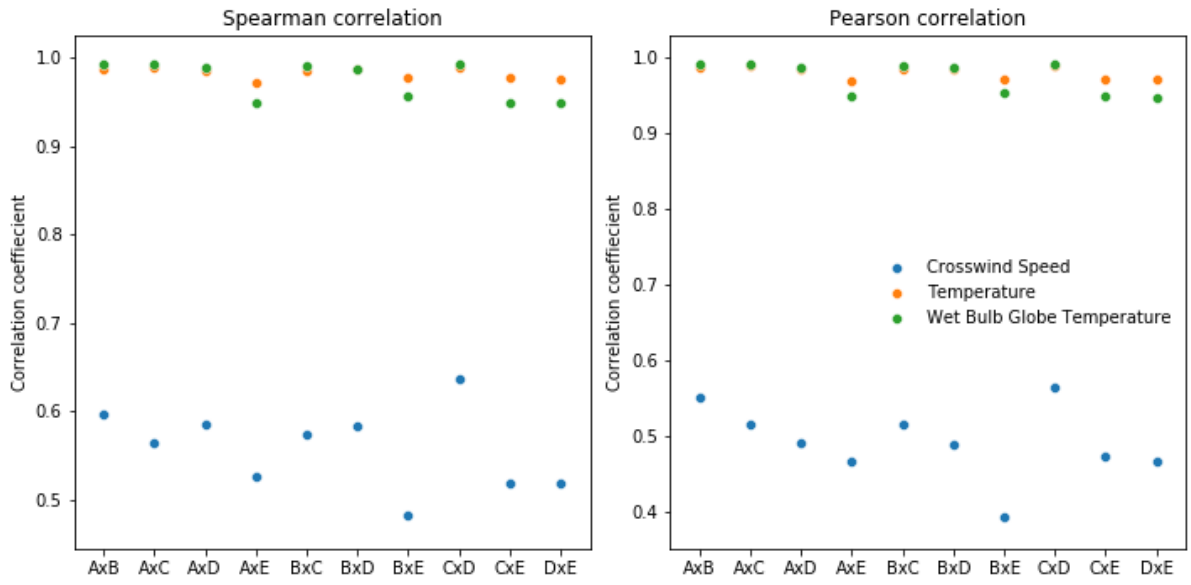


Figure 8: Scatterplot of Spearman and Pearson correlations between sensors dislayed on x-axis.

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

   When analyzing the correlations it can be observed that the correlations between sensors for the temperature and the wet bulb globe temperature are high, and all sensors correlate strongly with each other on these measurements, as the correlations are above 0.9 and close to 1.0 for both Spearman and Pearson correlations. However, the data from one sensor stands out, sensor E, all correlations between sensor E and another sensor are lower than the correlations between all other sensors.

   In contrast to the correlations between sensors for temperature and the wet bulb globe temperature, the correlations for crosswind speed are relatively low. Almost all correlations for crosswind speed are below 0.6, except for the spearman correlation between C and D at 0.63, the lowest correlations are, in correspondence with findings for temperature and wet bulb globe temperature with sensor E.

9. *If we told you that that the sensors are located as in figure 9, hypothesize which location would you assign to each sensor and reason your hypothesis using the correlations.*

Figure 9: Numbered locations of sensors on areal photo of Rijsenhout.

where will the sensors be located

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).*

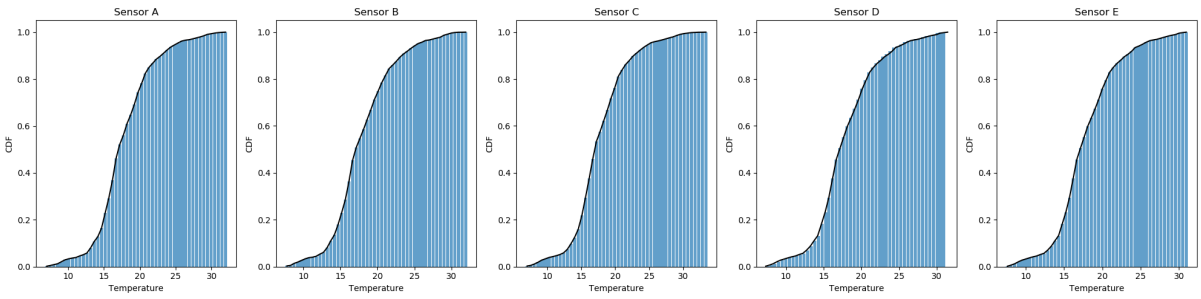


Figure 10: Cumulative Density Function of the temperature in °C per sensor.

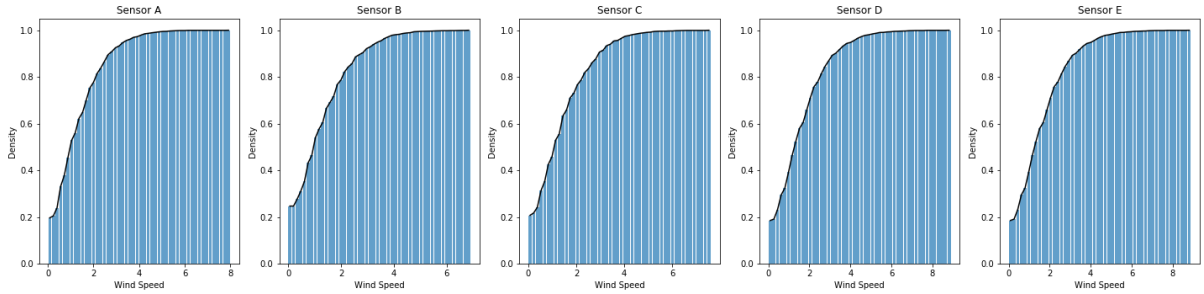


Figure 11: Cumulative Density Function of the wind speed in m/s per sensor.

The confidence intervals are calculated with the stats.t.interval function from Scipy. This function assumes the data has a student's t distribution, however as seen in question 6, the distribution for the wind speed is right skewed. This led to incorrect negative values for the lower limit of the confidence intervals for the wind speed each sensor, therefore the CDF graphs were used to determine the lower limit at a 95% confidence level. The plots in 11 show that about 20% of the values are 0, hence the value at 2.5% is 0, therefore this value is used as a lower limit in the confidence interval. The distributions for the temperature values are close to a student's t distribution, and only mildly skewed, hence the values for the confidence interval calculated via te stats.t.interval function do not need to be adjusted.

Table 4: Confidence intervals for temperature with 95% confidence level.

Sensor	*Lower limit*	*Upper limit*
A	10.160	25.778
B	10.071	26.060
C	10.045	25.781
D	10.128	25.864
E	9.799	26.909

Table 5: Confidence intervals for wind speed with a 95% confidence level.

Sensor	*Lower limit*	*Upper limit*
A	0	3.483
B	0	3.479
C	0	3.717
D	0	4.168
E	0	1.998

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

Table 6: t-statistic and p-values of the tested sensor pairs for temperature

Sensor pair	*t-statistic*	*p-value*
E-D	16.750	9.731×10^{-60}
D-C	6.875	7.843×10^{-12}
C-B	-10.581	1.274×10^{-25}
B-A	7.598	4.24×10^{-14}

Table 7: t-statistic and p-values of the tested sensor pairs for wind speed

Sensor pair	*t-statistic*	*p-value*
E-D	-44.709	1.839×10^{-320}
D-C	9.537	3.374×10^{-21}
C-B	6.069	1.486×10^{-9}
B-A	-2.485	0.013

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

    The hypothesis test are done with the stats.ttest_rel function from Scipy, as the samples are measures at the same location and are therefore not truly independent. The p-values for all hypothesis tests for the temperature and wind speed are below 0.025, and therefore should the null hypothesis that the tested sensors have identical means be rejected.

13. *Bonus:*
    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

    Code van bonus question en info

    ```
    If you have code put it here
    ```

References

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