GEO1001 - Assignment 1 Report

Ondrej Veselý

21st September 2020

Introduction

For this assignment we used the data-set containing climate data measured in 5 locations within the town of Rijsenhout [1]. The report is roughly subdivided into five sections related to each independent lesson (i.e. A1-4 + Bonus), with each section further split into a subsection for each corresponding question.

1 A1

1.1 Mean statistics

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

We observe fairly consistent mean statistics across all sensors for most of the variables. Of interest are Wind Speed, Crosswind Speed and Headwind Speed where Sensor E has significantly lower standard deviation and variance.

1.2 Temperature histograms

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?

See Figure 1. Number of bins that is too low doesn't give us a clear picture of the distribution. Higher numbers may introduce too much noise into the probability density estimation. Ideally we should choose the right amount of bins for our sample size using rule of thumb, e.g. Rice number (16 for our case).

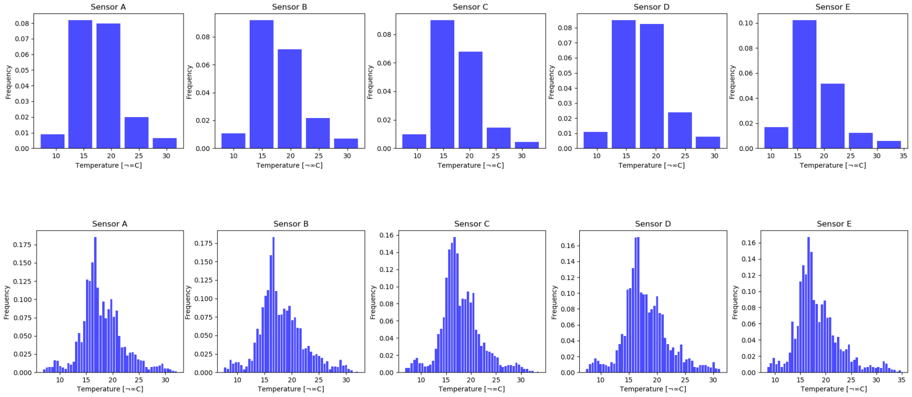

Figure 1: Comparison of different histogram binning values

1.3 Temperature frequency polygons

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

See Figure 2.

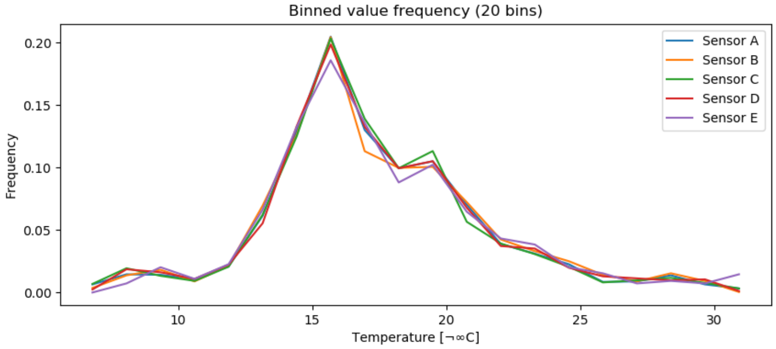

Figure 2: Temperature frequency polygons

1.4 Boxplots

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

See Figure 3

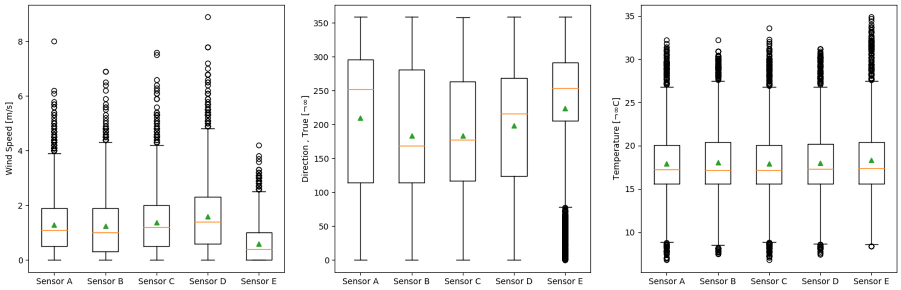

Figure 3: Boxplots of Wind Speed, Direction and Temperature

2 A2

2.1 Temperature distribution

Plot PMF, PDF and CDF for the 5 sensors Temperature values. Describe the behaviour of the distributions, are they all similar? What about their tails?

See Figure 4. The temperature measurement distributions are all fairly similar and could be all described as asymmetrical positively-skewed distributions. They roughly resemble normal distributions, although some anomalies in the tails can be observed.

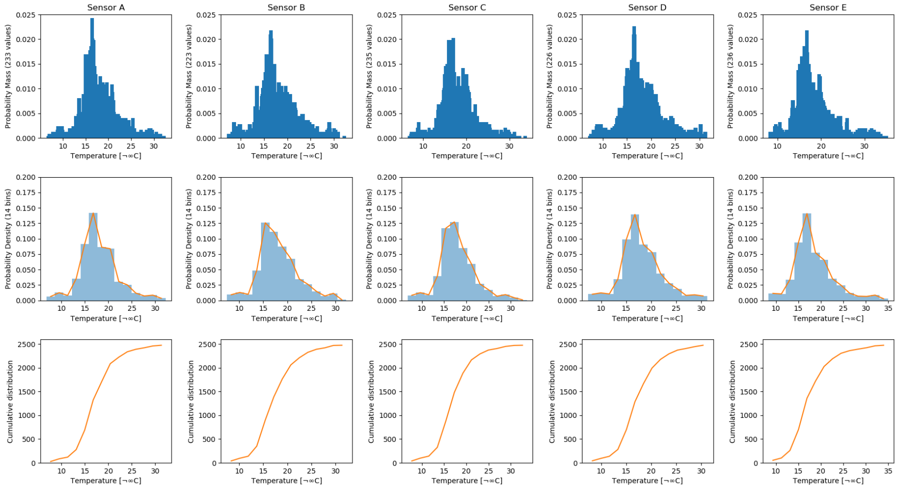

Figure 4: Probability mass, probability density and cumulative distribution functions of the temperature measurements

2.2 Kernel Density Estimation

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

See Figure 5. Assuming that PDF is smooth in reality, Kernel Density Estimation (KDE) smooths the PDF using a kernel function. The resulting estimated PDF shows less noise, introduced either by lower sample sizes, or high number of bins (i.e. in our case).

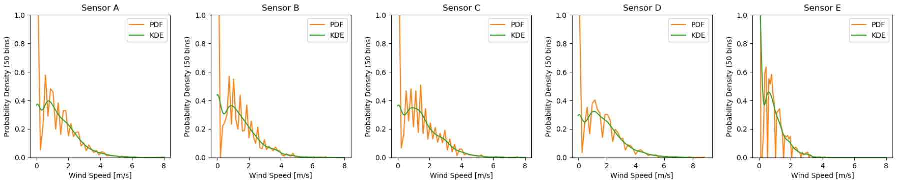

Figure 5: PDF and KDE for Wind Speed measurements (using 50 bins)

3 A3

3.1 Correlation scatter-plot

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.

See Figure 6

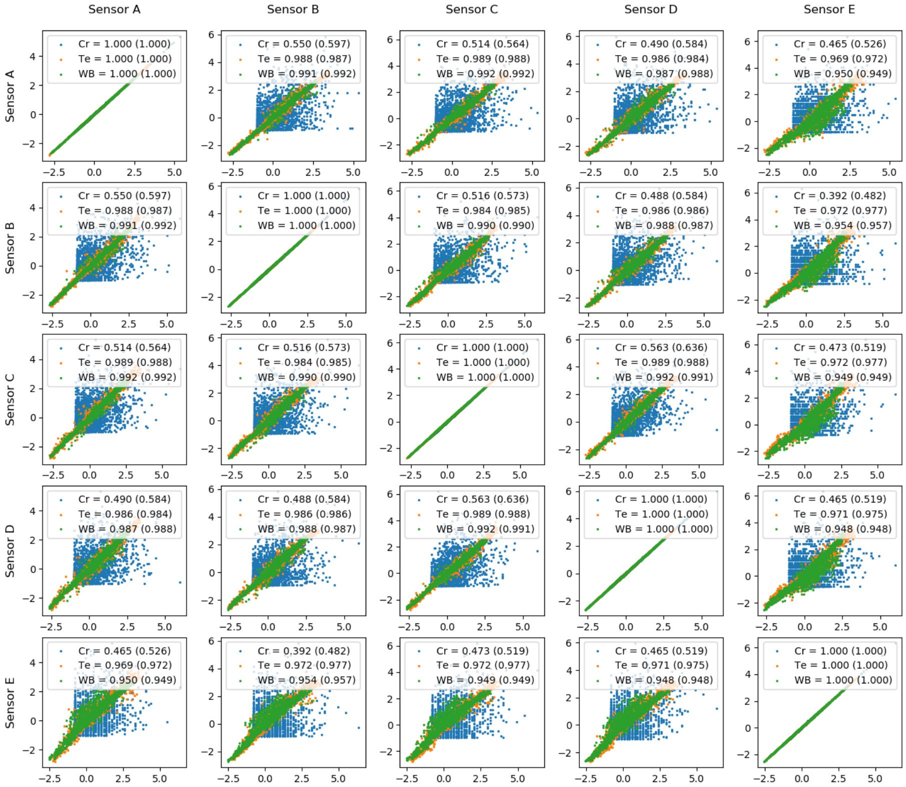

Figure 6: Pearson (and Spearman) correlation ranks between all sensors measurements of Temperature(Te), Crosswind Speed(Cr) and WBGT(WB)

3.2 Correlation conclusion

What can you say about the sensors' correlations?

We observe very strong ($r > 0.9$) correlations in Temperature and Wet Bulb Globe Temperature measurements in between all sensors. The correlations between Crosswind Speed measurements are weak or moderate ($0.4 < r < 0.6$).

3.3 Location hypothesis

If we told you that that the sensors are located as follows (Figure 7), hypothesize which location would you assign to each sensor and reason your hypothesis using the correlations.

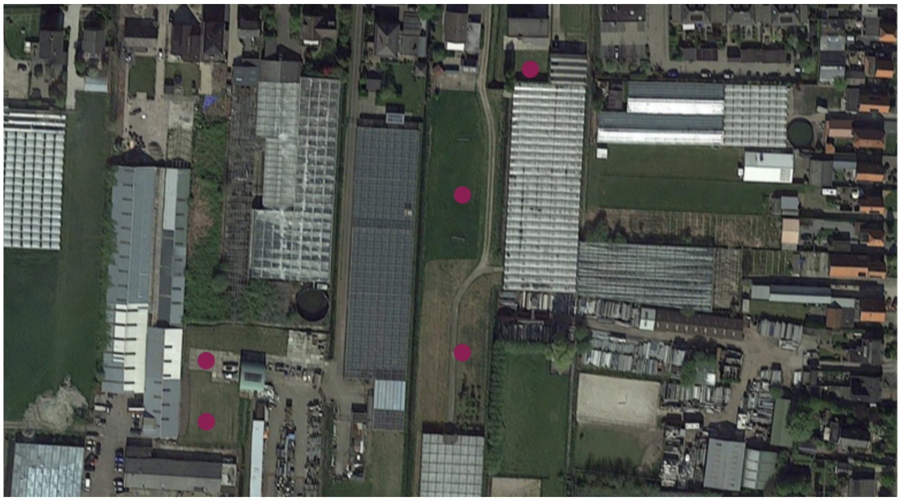

Figure 7: Situation map of the five sensors

Two pairs on sensors have been each situated in close proximity to each other, or in very similar conditions. One sensor has been placed in a separate location. According to our hypothesis we should identify the two pairs by their stronger correlations and the one 'outlier' sensor by its weaker correlation to the rest of the group.

The two mutually exclusive pairs with strongest correlations are:
Sensor C, D *Crosswind speed 0.563, Temperature 0.989, WBGT 0.992*
Sensor A, B *Crosswind speed 0.550, Temperature 0.987, WBGT 0.991*

The remaining sensor E consistently has the lowest correlation with the rest of the sensors:
Sensor E, x *Crosswind speed <0.473, Temperature <0.972, WBGT <0.954*

4 A4

4.1 CDF and confidence intervals

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).

See Figure 8 and Table 4.1

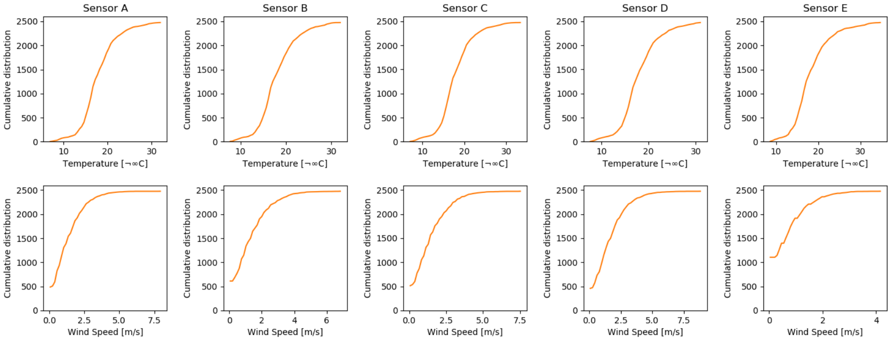

Figure 8: Cumulative distribution functions of Temperature and Wind Speed

	Sensor A	Sensor B	Sensor C	Sensor D	Sensor E
Temperature	17.81 - 18.13	17.90 - 18.23	17.75 - 18.07	17.84 - 18.15	18.18 - 18.53
Wind Speed	1.25 - 1.33	1.20 - 1.29	1.32 - 1.42	1.53 - 1.63	0.57 - 0.62

Table 1: Temperature and Wind Speed 95% confidence intervals

4.2 Similarity hypothesis

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*

See Table 4.2

	E, D	D, C	C, B	B, A
Temperature	t=3.000 p=0.003	t=0.729 p=0.466	t=-1.324 p=0.185	t=0.841 p=0.400
Wind Speed	t=-32.673 p=0.000	t=5.871 p=0.000	t=3.893 p=0.000	t=-1.501 p=0.134

Table 2: T and P values for tested sensors

4.3 CDF and confidence intervals

What could you conclude from the p-values?

We observe statistically significant differences in Temperature time series for sensor-pair E,D and for Wind Speed time series for sensor-pairs E,D; D,C and C,B, which means that the measured differences unlikely occurred by chance.

5 Bonus

5.1 AC energy demand

Your "employer" wants to estimate the day of maximum and minimum potential energy consumption due to air conditioning (AC) 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.

For the code see accompanying Python Notebook.

The 3 hottest days in time-series are:

1. Friday 26 June 2020
2. Thursday 25 June 2020
3. Wednesday 24 June 2020

The 3 coolest days in time-series are:

1. Tuesday 14 July 2020
2. Wednesday 10 June 2020
3. Wednesday 08 July 2020

Further more, we consider, that the AC could be set to turn on only in temperatures above certain threshold. I.e. if we turn on AC only in temperatures over 25C, the 3 most energy demanding days in time-series would be:

1. Wednesday 24 June 2020
2. Thursday 25 June 2020
3. Friday 26 June 2020

References

[1] Daniela Maiullari and Clara Garcia Sanchez. Measured climate data in rijsenhout, 2020.