Assignment 01

Statistical analysis of a heat stress measurement dataset

Deadline is **21st September 2020 by 24h**.

Late submission? 10% will be removed for each day that you are late.

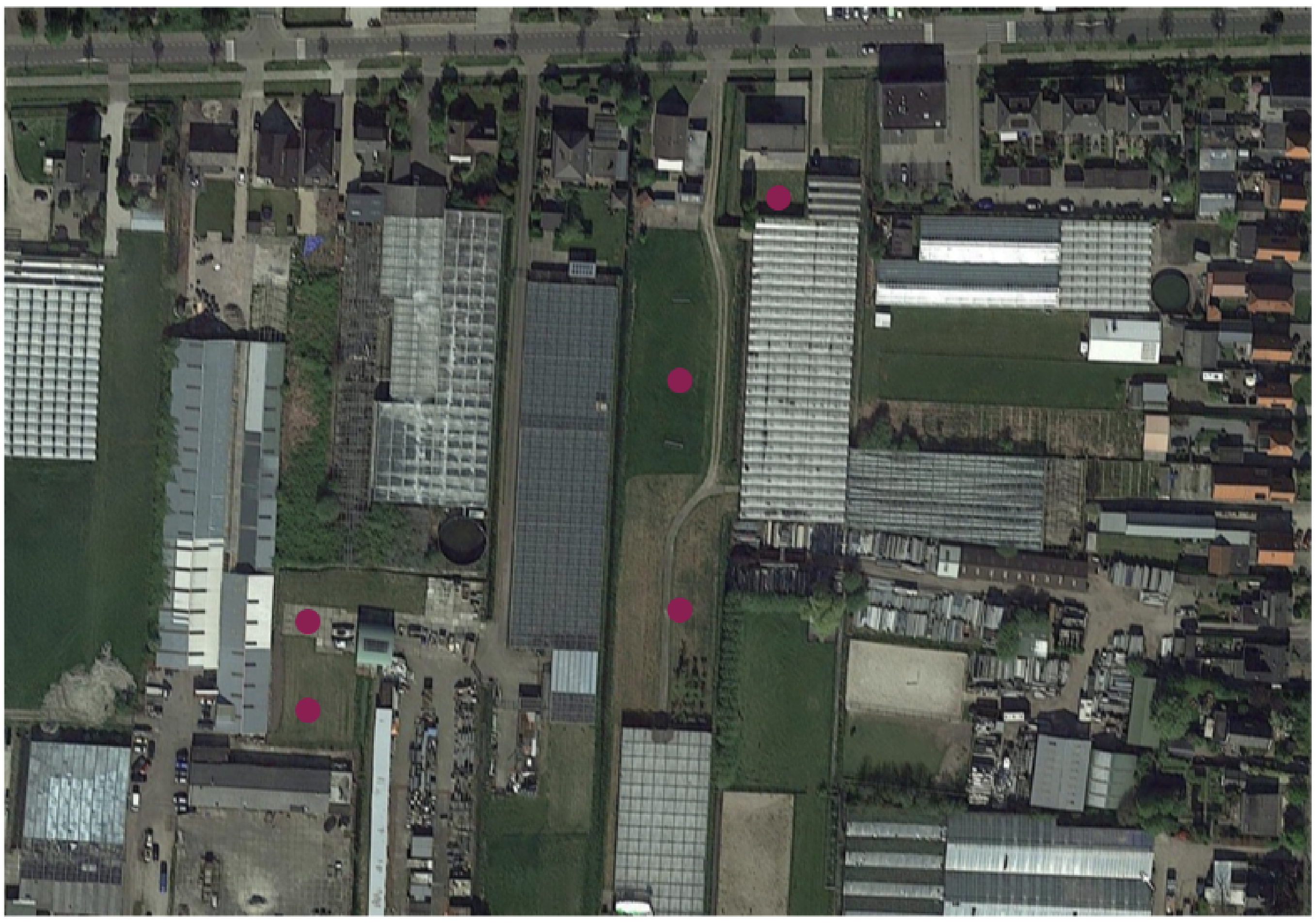
Overview

For this assignment you will use data collected from 5 heat stress sensors placed somewhere in the Netherlands during this summer. The sensors are Kestrel 5400 and their specs are included within the assignment materials. In order to identify if the dataset is of any value to your "employer", it is your job to deeply analyse the dataset and derive hypothesis from it.

The work you need to do for this assignment can roughly be subdivided in 4 tasks related to each independent lesson:

1. After lesson A1:
   - Compute mean statistics (mean, variance and standard deviation for each of the sensors variables), what do you observe from the results?
   - 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?
   - Create 1 plot where frequency poligons for the 5 sensors Temperature values overlap in different colors with a legend.
   - Generate 3 plots that include the 5 sensors boxplot for: Wind Speed, Wind Direction and Temperature.
2. After lesson A2:
   - 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?
   - For the Wind Speed values, plot the pdf and the kernel density estimation. Comment the differences.
3. After lesson A3:
   - 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.
   - What can you say about the sensors' correlations?

- 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.



4. After lesson A4:

    - Plot the CDF for all the sensors and all the variables, then compute the 95% confidence intervals for all the variables and sensors and save them in a table (txt or csv form).
    - 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?

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.

To complete the exercises we recommend you to use the already available functions from libraries in python (such as scipy.stats or pandas). Although, if you prefer it, you can also code the functions yourself.

What we give you

hw01.tar.gz

This zip file contains 5 CSV files correponding to 5 different heat stress sensors placed within a neighbourhood in the Netherlands. In addition there is a readme file that further explains the diverse variables and a document specifying the accuracy of the measurements.

Input assumptions

- You should use all the time series measurement data unless is specified otherwise in parts of the exercise.
- You may process variables that have missing data, since this is common issue in measurement collection.

Deliverables

Each individual student has to submit 2 items:

1. Your code in a Github repository with a README file explaining its use for each of the 4 parts within this assignment. You can have multiple python files (like 4, 1 for each part) or just one main file if you prefer, as far as you always have one main python file that executes all your code, so we can verify it works.
2. Report in PDF format where only figures and your answer to the questions are reasoned. The PDF has to be compiled with latex, and you should also include the latex source code in your repository. If you prefer you can also share an overleaf folder. The data should be cited within the report bibliography using the DOI: 10.4121/12833918.v1.

Code specifications

Your Python source file(s) need to have the following text at the very top:

```
#-- GEO1001.2020--hw01
#-- [YOUR NAME]
#-- [YOUR STUDENT NUMBER]
```

The main file to run your code should be named `geo1001_hw01.py`. I should thus be able to execute the code by calling the main file like this:

```
python geo1001_hw01.py
```

Apart from that you are free to organise your code to your liking.

The report to submit

You need to submit a report explaining briefly the conclusions from the plots you created with respect to the data. In addition, you should address each of the questions included within the description of the assignment.

We expect maximum 1500 words for this.

Marking

	Marks
followed all specifications and runs without modifications	1
report quality/clarity	0.5
lesson A1	2
lesson A2	2
lesson A3	2
lesson A4	2
bonus question	0.5

Submission

- Do *not* submit your assignment by email, but upload the pdf file for the report in the following surfdrive folder. You need to upload in addition **one zip file** with the source latex code. You can also share this through Github if you prefer.
- The code source files in python need to be shared through Github.
- The name of the zip/pdf files should contain your family names.