Modified preregistration template for Data Mastery Challenge course

Source of original: OSF

Study Information

1. **Title**
   Urban green spaces and variation in cooling during warmer summer months in the moderate maritime climate zone: the case of UT Campus Enschede

2. **Authors**
   J.K. Powers
   J.W. Njoroge
   W.H.Jansen

3. **Description**
   The original research is done in Paramaribo, the capital of Suriname. Suriname is a former Dutch colony in the Caribbean (north of Brazil). We had a series of projects on urban green in Paramaribo (which included engaging citizens to monitor sensors of urban climate in the city.

   The data derived from climate sensors was analyzed to understand the effect of urban green on the tropical urban climate. The existing analysis was published as a scientific article: L. Best, N. Schwarz, D. Obergh, A.J. Teuling, R. van Kanten, L. Willemen (2023): Urban green spaces and variation in cooling in the humid tropics: The case of Paramaribo. Urban Forestry and Urban Greening

   The goal of our research is to analyze the effects of urban green spaces on cooling in the moderate maritime climate zone found in Enshede, NL. We will transfer the analysis techniques from the Paramaribo study to analyze data collected from sensors located on the UT campus, thus validating the research that has been previously done.

4. **Hypotheses**

- If urban green spaces affect temperature in warmer summer months in moderate maritime climate, the mean temperature will be lower in areas that have more greenery and higher in areas that have no greenery on the UT Campus. (directional hypothesis)
- If urban green spaces affect temperature in cooler winter months in moderate maritime climates, the mean temperature will be cooler in areas that have more greenery and lower in areas that have no greenery on the UT Campus. (directional hypothesis)

Design Plan

5. Study type (required; changed from original template)
    - Replication: The environment around the sensors in Paramaribo were analyzed and classified regarding vegetation.

      For parallel research the environment surrounding the sensors in Enschede also should be analyzed and classified in the same manner. We will use a landuse map from the University of Twente campus management to classify the environment around the sensors, and we will visit each sensor to verify the land surface type directly under the sensors.
    - Replication: The produced data in Enschede should have the same parameters as the parameters in Paramaribo.
    - Adaption: The original research focused on humid climate. For the purposes of our study in the Netherlands we will focus moderate maritime climate which has warmer summers and windy colder winter months. The original data visualization will be adapted to describe the new scenario.

      In this case the relevant situation is going to focus on: Urban green spaces effect on moderate maritime climate zones

6. Study design (required)
    - **Data Processing and Collection:**
      Temperature and humidity data will be analyzed over a four-year period (9 Jan 2021 to 22 August 2024) and focus on a comparison between summer months (June, July, and August) and winter months (December, January, and February). The locations of the sensors that collect the local temperatures will be documented and a description of their surroundings will be created. The sensor data will be validated and checked for inconsistencies. All sensor data will be compared without concern for whether it is in a green space or urban space to remove bias.

      Land surface temperatures will be calculated over a range of 5 years (1 Jan 2019 – 31 December 2024) using data from Landsat 8/9 OLI/TIRS images from

EarthExplorer (https://earthexplorer.usgs.gov/). The images will be pre-processed in QGIS and clouds will be masked using CloudMasking plugin. Land surface temperatures will be calculated based on the approach followed in the Paramibo study which includes atmospheric corrections, calculating brightness temperature, DVI and surface emissivity.

We will use a landuse map from the University of Twente campus management to classify surface type around the sensors, and we will visit each sensor to verify the land surface type directly under the sensors. Surface cover will be designated as either “no vegetation”, “grass and shrubs”, “trees only”, or “trees over shrubs”.

Methods:

CODE

Objective is to translate the code from R(from the original Paramaribo study) to Python with the purpose of doing the following data analysis:

- Step 1: Calculate the difference in temperature between each location and the average of all locations
- Step 2: Compare the temperatures of all locations to each other at sunrise and sunset for summer (6:00hrs and 22:30hrs) and for winter (9:00hrs and 17:00hrs)
- Step 3: Compare the night-time temperatures for extremes between 23:30hrs and 6:00hrs for summer and 18:00hrs and 9:00hrs for winter
- Step 4: Calculate aggregate micro-climate indicators: Mean annual and seasonal temperature and humidity, minimum local night-time temperature over the year and seasonally, range of night-time temperatures, hot extremes or max temperatures and humidity on the ten hottest days identified
- Step 5: Estimate share of land cover via visual interpretation within two buffer zones around sensors at 10m and 300m. For the 10m buffer zone, 40 randomly generated sampling points will be used with a minimum distance of 2m, and for the 300m buffer zone, 50 randomly generated sampling points will be used with a minimum distance of 10m.(optimization depending on time)
- Step 6: Analyze the relationships between location characteristics and micro-climate indicators to avoid collinearity and reduce the number of variables to consider using Pearson correlations to exclude numerical variables.
- Step 7: Investigate mosaic plots and box plots to identify close relations.
- Step 8: Investigate relationships between location characteristics and micro-climate variables using Pearson correlations for numberic and Kruskal-Walls test for categorical location characteristic
- Step 9: Adjust for multiple comparisons using steps followed in the Paramibo study and use the Hommel adjustment for the correlations and the Wilcoxon tests.

Sampling Plan

7. Existing data
    - Sensor Data (hourly)
    - Landsat satellite imagery
8. Explanation of existing data (optional)
    - Sensor Data is collected from weather sensors located on UT Campus. This data will provide precise local temperatures.
    - Landsat satellite imagery will provide land surface temperature.
9. Data collection procedures (required)
    - Site data will be collected through field work, OpenStreetMap, and Maxaar satellite image in the Bing repository in QGIS software using random sample points. Surface cover will be designated as either “no vegetation”, “grass and shrubs”, “trees only”, or “trees over shrubs”.

Variables

10. Manipulated variables (optional)
11. Measured variables (required)
    - Temperature [C°] and Humidity [%] -
    - Temperature and humidity data will be from over a five-year period (1 Jan 2019 to 31 December 24) and focuses on the summer months (June, July, and August) and winter months (December, January, and February).
    - Environmental descriptors: “No vegetation”, “Grass and Shrubs”, “Trees only”, and “Trees and Shrubs”
    - Land Surface Temperature
        - Land surface temperature over a range of 5 years (1 Jan 2019 – 31 December 2024) is from Landsat 8/9 OLI/TIRS images from EarthExplorer (https://earthexplorer.usgs.gov/).
    - Dominant surface cover (10m buffer)
    - Dominant vegetation (10m buffer)
    - Dominant surface cover under sensor
    - Elevation
    - Distance to campus center (Vrihof)
    - Impervious in 10m buffer
    - Total green in 10m buffer
    - Trees in 10m buffer
    - Impervious in 300m buffer
    - Total green in 300m buffer
    - Trees in 300m buffer
12. Indices (optional)
    - Daily temperature and humidity data will be analyzed at all locations for summer (6:00hrs and 22:30hrs) and for winter (9:00hrs and 17:00hrs) to calculate the following:

    - Max temperature in hot extremes
    - Max humidity in hot extremes
    - Average temperature yearly
    - Average temperature in core summer months
    - Average temperature in core winter months
    - Average humidity yearly
    - Average humidity in core summer months
    - Average humidity in core winter months
    - Highest temperature mean during the day
    - Lowest Temperature mean during the day
    - Lowest Standard deviation temperature during the day
    - Highest humidity mean during the day
    - Lowest humidity mean during the day
    - Highest mean standard deviation of humidity during the day
    - Lowest mean standard deviation humidity during the day
    - Average nighttime temperature yearly
    - Average nighttime temperate for core summer months
    - Average nighttime temperature for core winter months
    - Nighttime temperature range for summer
    - Nighttime temperature range for winter
    - Nighttime min temperature for summer
    - Nighttime min temperature for winter
    - Nighttime max temperature for summer
    - Nighttime max temperature for winter
- Land surface temperature will be averaged and standardized to compare periods for summer and winter.

13. Analysis Plan

    Our analysis plan includes the following steps:

    - Explore and analyze the relationships between location characteristics and micro-climate indicators to avoid collinearity and reduce the number of variables to consider using Pearson correlations to exclude numerical variables.
    - Investigate mosaic plots and box plots to identify close relations.
    - Investigate relationships between location characteristics and micro-climate variables using Pearson correlations for numeric and Kruskal-Walls test for categorical location characteristics.
    - Adjust for multiple comparisons and use the Hommel adjustment for the correlations and the Wilcoxon rank sum tests.

14. Statistical models (required)

    - We will use Pearson Correlation with Hommel adjustment (for multiple comparisons) to exclude numerical variables out of a pair of variables with significant correlation coefficient > 0.7 and to analyze the share of land cover.

    - Predictors: “No vegetation”, “Grasses and Shrubs”, “Trees Only”, and “Trees over Bushes”
      These variables have characteristics that are analyzed and include: Elevation, distance to water, distance to CBD, share of total green in 10m buffer, share of trees in 10m buffer, share of impervious in 10m buffer share of total green in 300m buffer, share of trees in 300m buffer, share of impervious in 300m buffer
    - Exploratory process to identify share of land cover
- We will use Kruskal-Wallis using pairwise Wilcoxon rank sum tests for categorical location characteristics and comparing between groups
    - Predictors: “No vegetation”, “Grasses and Shrubs”, “Trees Only”, and “Trees over Bushes”
    - Exploratory process to identify share of land cover
- We will run a linear regression with significant variables discovered during our exploratory process to see if there are any statistically significant correlations between the variables.
- **More information**: This is perhaps the most important and most complicated question within the preregistration. As with all of the other questions, the key is to provide a specific recipe for analyzing the collected data. Ask yourself: is enough detail provided to run the same analysis again with the information provided by the user? Be aware for instances where the statistical models appear specific, but actually leave openings for the precise test. See the following examples:
    - If someone specifies a 2x3 ANOVA with both factors within subjects, there is still flexibility with the various types of ANOVAs that could be run. Either a repeated measures ANOVA (RMANOVA) or a multivariate ANOVA (MANOVA) could be used for that design, which are two different tests.
    - If you are going to perform a sequential analysis and check after 50, 100, and 150 samples, you must also specify the p-values you’ll test against at those three points.

15. Transformations (optional)
16. Inference criteria (optional)
17. Data exclusion (optional)
    Datapoint with missing information will be excluded out of the research.
18. Missing data (optional)
19. Exploratory analysis (optional)

There is potential to compare findings with the original study.

Other

20. Other (Optional)