Forest Information Technology
Environmental Data Analysis
Final exam

Luis Miranda
luis.miranda@hnee.de

July 2021

1 General Information

1. The final exam of the module *Environmental Data Analysis* is a project report. In the Summer Semester 2022 this report refers to work on height models, as measured with laser. The report is due officially on July 29, 2022. Submissions will be admitted until August 15, 2022.
2. This document and the data needed for the exam are available on the Moodle platform of the course, at `https://lms.hnee.de/course/view.php?id=532`
3. Please submit your report and any additional materials that you consider useful as a single file (e.g. ZIP) either over the Moodle platform or via email to luis.miranda@jnee.de
4. Not only the implementation of the analyses will be graded, but also the overall context, interpretation and description of the problem in your report.

2 Data description

1. The file `Mehrfachdownload_rIYYX7_2VX73u.zip` contains laser measurements from 20 tiles in the German state of Thuringia. The tiles are 1x1km in size. Each of the 20 files has also a metadata file.
2. The data is released for public use and was gathered fromn the Geoportal of the Land Thuringia.
3. You can find more detailed information, including shapefiles here.
4. Each file contains coordinates and a single column with the corresponding height values.

3 Task description

1. **Metrics.** For each data file, calculate the following statistic metrics of the height:
   - Mean height
   - Standard deviation
   - Kurtosis
   - Skewness
   - Maximum and minimum heights
   - Range (maximum-minimum)
   - Interquartile range
   - All percentiles from 5 to 95 in increments of 5 (5, 10, 15, ..., 90, 95), for a total of 19 percentile values
2. Create a table with the 27 metrics and the 20 tile identifiers. The table should be of dimensions 20x27.
3. Optionally, you can explore the data with a boxplot (20 boxplots) to have a first impression of important differences in the data between tiles. See Fig. 1 for reference.
4. Optionally, localise the tiles in a map.
5. **PCA.** Perform a Principal Component Analysis over the metrics to explore the latent dimensionality of the dataset. Create and interpret a scree plot and biplots of PC1-PC2; PC2-PC3 and PC1-PC3. Interpret your results. Take care to exclude the tile identifiers from your analysis.
6. Select a subset of the principal components carrying the most variance. How many PC-variables can be used? Justify your decision.
7. Which are the most important metrics for the PC-variables that you selected?
8. **Clustering.** Perform an unsupervised clustering analysis with the PC-variables you selected. Try different numbers of clusters including 2, 3 and 4 groups.
9. If you decided to make boxplots with the original data, compare the results of the clustering with those of the exploratory plots.

10. Do the groups created by the unsupervised algorithm agree with the reality? Can you make some sense of the clustering?

11. **Classifier.** For these tasks, use the same PC-variables that you selected as most important.

12. Implement a decision tree classifier to classify the 20 tiles into the number of groups that you selected in the previous section. Generate a plot of the classifier and give your interpretation.

13. Implement a random forest classifier to repeat the classification. Since the data table is very small, you need to take very much care with the tuning of the hyperparameters of the trees (number of trees, depth, pruning, etc).

14. Save your classifier to be tested in the next section.

15. **Classification workflow.** Select at least 2 tiles from the metrics table and run a classification:

    a) From the metrics calculate the 27 PC-variables. Select the most important ones as per your analysis.

    b) Use them as inputs to the random forest classifier and verify that you get a predicted class for each tile.

    c) Interpret the classification results.

    → This step should be done with a **test** dataset that was not involved in the model building. However, this would require more tiles to be downloaded and included. Alternatively, we could have built the models with less (e.g. 18) tiles, but those are already very few data points. Be aware of this limitation.

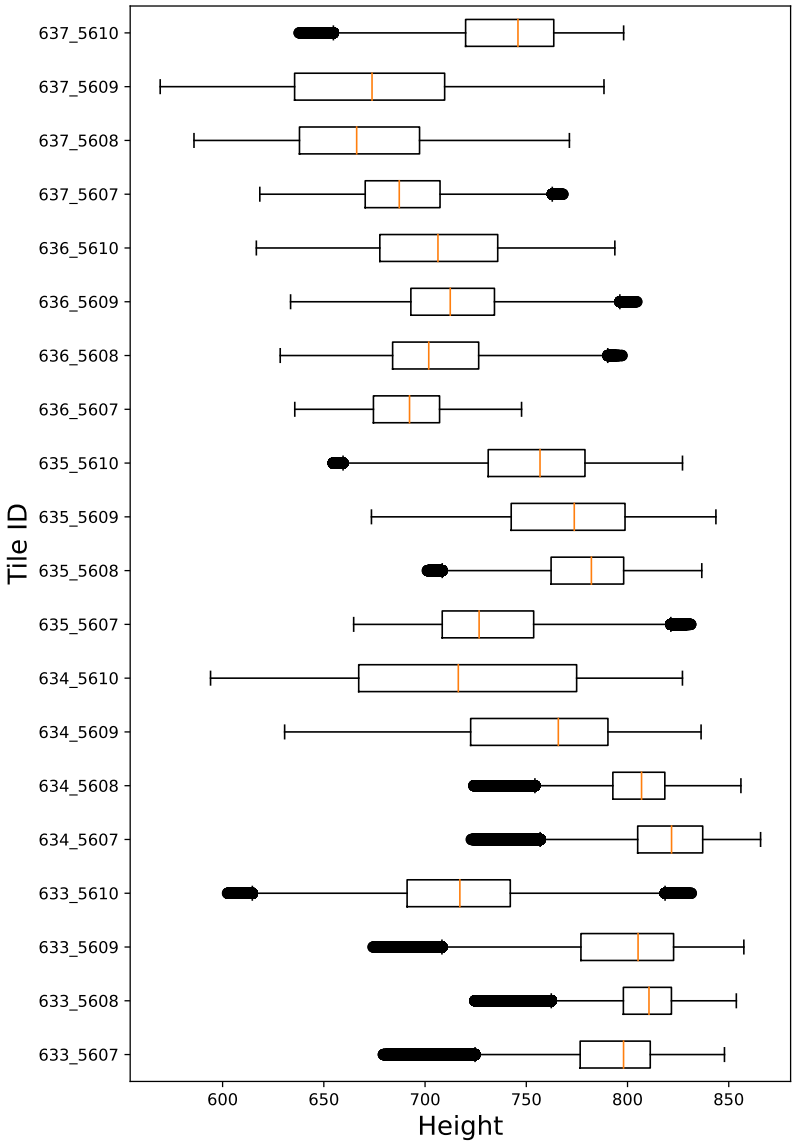

Figure 1: Exploratory representation of the height data in 20 tiles in German Thuringia, from airborne laser measurements.