Project B4: Effects of invasive species on native biodiversity

Reelika Pärnpuu
Marta Miia Pärnpuu
2023

Link to repository: carvin-to-go/AT_invasiivsed: Data Science project analysing effects of invasive species on native flora (github.com)

Task 2. Business understanding

Background. Our university's plant ecology workgroup seeks a better understanding of how invasive species affect native biodiversity. The workgroup itself will mostly focus on the connection between soil parameters and changes in native biodiversity, but we will look for more general trends in the provided data. We will report our findings to the plant ecology workgroup, and they will, when the project ends, report to the Republic of Estonia Environmental Board.

Goals. The general goal is to prevent the spread of invasive species, mostly by analysing and implementing new invasive species control methods. We will be working on data from before the control methods have been implemented, but the project will go on to analyse the same areas after implementation. A narrower goal would be to find ways for the plant ecology workgroup to utilise Python in their work since they usually use R packages.

Success criteria. Success criteria aren't quite relevant here. Hopefully, the invasive species control methods can be implemented; if not, we will take into consideration that failures are a big part of the scientific process.

Inventory of resources. Our resources for this project are our team of two, people from the plant ecology group who can provide general information to us, if necessary. We also have three datasets, the first of which is the geographical locations of the sampled areas, the second is the list of plants found in those areas (floral biodiversity), and the third dataset is the results from the soil analysis performed on the sample areas. Our software resources include mostly Jupyter Notebook and Microsoft Excel.

Requirements, assumptions, and constraints. Requirements for our work are looking for general trends in the data provided and, if time permits, calculating the biotic novelty

measure. Time constraint for our work to be finished by the 14th of December, there are not any constraints besides this.

Risks and contingencies. There are not any significant risks for this project. The only relevant risk, which cannot be easily dealt with, would be either of the team members getting sick, in which case discussions should be had with the course instructors.

Terminology. This project will use the following terminology:

Firstly, for our purposes, biotic novelty is a specific measure which helps analyse functional differences between native and non-native introduced species and temporal dynamics of species introductions; this will be calculated using predefined formulas from Schittko, C, Bernard-Verdier, M, Heger, T, et al. A multidimensional framework for measuring biotic novelty: How novel is a community? Glob Change Biol. 2020; 26: 4401–4417. (https://doi.org/10.1111/gcb.15140).

Secondly, floral biodiversity data has been gathered using the quadrat method, which in this case means that a quadrat (small square frame) has been placed on the ground, and the percentage cover of the species has been reported.

Lastly, sampled areas for our purposes mean the 24 areas from which data was gathered. Since each area had three quadrats placed in it (sub-sample areas), for floral biodiversity purposes, it's essential to keep these two terms separate.

Costs and benefits. Costs and benefits are not relevant here.

Data-mining goals. Our data-mining goals include the analysis of three datasets, mainly focusing on PCA analysis of the connection between floral species and pH measurements. Calculating the biotic novelty measure.

Data-mining success criteria. Success criteria are not very relevant here, we will mainly focus on completing our goals to the best of our ability.

Task 3. Data understanding

Data requirements. To achieve our goals, we need data about sampled areas - for example, what plants there are and the soil chemistry. We received the necessary data files from our university's plant ecology workgroup.

Data availability. We have access to three Excel files with the names: "EEB_samples.xlsx", "EEB_soil_chemistry.xlsx", and "Plant_survey_2023_final.xlsx". The Excel files open, and the data seems to be formatted correctly. We also tried to read these files in Jupyter Notebook with the pandas module, this was successful. Therefore the data is in usable shape.

Selection criteria. Firstly, from the file **"EEB_samples.xlsx"** we will use the sheet with the name **"full_table"**; sheets "solidago" and "reynoutria" aren't necessary because the information is included in the sheet "full_table". Table "full_table" (shape 216 x 14) contains columns: 'new_code', 'project', 'plant_sp', 'sampling_area(on_bag)', 'treatment', 'sampling_area_nr', 'gradient', 'grad_nr', 'plant_nr', 'expert', 'area', 'lat', 'lon', 'status _dna_extraction'. Column 'expert' is not necessary. Column "project" can also be dropped since it contains only one value "EEB". Column 'status _dna_extraction' isn't relevant to our work.

Secondly, **"EEB_soil_chemistry.xlsx"** contains soil chemistry sample analysis from every location, we will be using the column "pHKCl".

Lastly, we will use all columns from the file **"Plant_survey_2023_final.xlsx"**, which contains floral biodiversity data: sample codes and the counts of different plant species in the sample area.

Describing data. Firstly, in the **"EEB_samples.xlsx"** data frame, there are a total of 216 sub-samples. The samples have been taken from 24 locations; in each location the quadrat method has been used to collect data from 3 sub-sample areas. For each quadrat, the sample was taken 3 times.

For the **"EEB_samples.xlsx"** data frame, we are primarily interested in the column "gradient", which has 3 categorical values "invaded", "transitional" and "natural". "Invaded" in this case, means that the invasive species are frequent in the sample area; "natural" means that the sample area only has local fauna; "transitional" sample areas are those in between the natural and invaded areas.

Secondly, **“EEB_soil_chemistry.xlsx”** contains a table (shape 25 x 11) with columns: 'ala_kood', 'Proovi nimetus', 'Nr.', 'Jrk. nr.', 'pHKCl', 'P-AL', 'K-AL', 'Ca-AL', 'Mg-AL', 'N -Kjeldahl', 'C-Tjurin'. It has information about the area where the sample was taken from, the pH of the soil, how much of different compounds (P-AL, K-AL, Ca-AL, Mg-AL) are in the soil in units milligram per kilogram. 'N -Kjeldahl' and 'C-Tjurin' are percentages of nitrogen and carbon content, respectively. Kjeldahl and Tjurin denote the names of methods. In the Jupyter Notebook, there is an additional row with units, which will have to be excluded while using the data for analysis. The table contains one soil chemistry sample from every location (there are 24 locations).

Thirdly, **“Plant_survey_2023_final.xlsx”** contains a table (shape 216 x 222) with sample codes and the counts of different plant species in the sample area.

Exploring data. The sub-sample areas are not quite evenly distributed. There are 63 samples from the west-Estonia, denoted as “W” in column ‘area’; 54 from the east-Estonia (“‘E”); 54 from the south-Estonia (“S”) and 45 from the north-Estonia (“N”). Other than this, there are no other issues. There are 72 “invaded” data points, 72 “natural” and 72 “transitional” data points, so the data is, in that sense, balanced.

Verifying data quality. Data quality seems quite good. The only thing we discovered was that in the ‘Antsla’ sampling area, the coordinates (latitude and longitude) were in red color. This could indicate that these are not accurate, but it should not be a big problem.

Task 4. Planning your project

Tasks:

1) Reading in the data, cleaning data, selecting relevant data, combining data frames [Reelika (4h), Marta (1h)]
2) Reading the article, understanding article [Reelika (3h), Marta (4h)]
3) Calculating BNI (index of biotic novelty), contains 7 steps [Reelika (6h), Marta (5h)]
4) Doing PCA [Reelika (5h), Marta (2h)]
5) Geographic correlation [Reelika (4h), Marta (4h)]
6) Analysing results/data, making graphs [Reelika (6h), Marta (4h)]
7) Making poster [Reelika (2h), Marta (10h)]

List of methods and tools:

1) Jupyter Notebook
2) BNI calculation. There are seven steps to calculate the BNI: (1) obtaining a trait matrix, (2) converting the trait matrix into a distance matrix, (3) obtaining species' first records, (4) converting the first records into a temporal coexistence matrix, (5) weighting the distance matrix by the temporal coexistence matrix, (6) multiplying the distance matrix by the species' relative abundance (optional), and (7) calculating the sum of all pairwise comparisons from the distance matrix.

$$\mathrm{BNI} = \sum_{i=1}^{s-1} \sum_{j=i+1}^{s-1} d_{ij} \times c_{ij} \times p_i p_j,$$

3) For PCA we will use sklearn.decomposition
4) Plots - seaborn and matplotlib
5) For poster we will probably use Canvas