Globally endangered species vulnerability to climate change in Europe

Team: Argo Ronk

Link of the repository:
https://github.com/argoronk/Climate_Change

Task 2. Business understanding

- Identifying your business goals:

Background: Current changes in climate and intensifying human activities, are the main drivers responsible for the decline in global biodiversity. Understanding how globally endangered species perform under climate change is fundamental for designing effective conservation strategies, ensuring the persistence of biodiversity, and promoting ecosystem resilience in the face of ongoing environmental changes.

Business goals: Determine the current distribution of globally endangered species in Europe and using climatic (bioclimatic) variables (future projections) to determine how and if the distribution of globally endangered species in Europe could change.

Business success criteria: Globally endangered species distributions in Europe are determined and mapped both for nowadays and for the future. Potential distribution changes (shifts) under current and future climates have been determined and visualized.

- Assessing your situation

Inventory of resources: This project will be carried out by one person (Argo Ronk). Data for this project needed (species occurrence data, current and future climate projections, essential map layers) is all open access and available online. Personal computer will be used as primary hardware for carrying out the analyzes and necessary software (Python with required libraries, ArcGIS) is already installed on it.

Requirements, assumptions, and constraints: As the project have to be finished by latest 11th of December (constrain), it is foreseen that first required data collection is finished by 20th of November. Required analyzes done by 5th of December. Results visualized by 8th of December and poster ready for submission in 10th of December.

Risks and contingencies: Risk 1 – Data may not be available (species occurrence data, current and future climates projections) due to technical reasons. It may be possible to use an alternative data source (there could be alternative download sources, data may be published as supplementary material in scientific articles). Risk 2 – Number of species to analyze is too large

to finish project in time. To mitigate this risk, one could take only subset of species (based on some threshold) into analyzes to ensure timely completion of the project. Risk 3 – staff could get sick as project is carried out during winter season. This risk can't be fully mitigated as project is carried out by one person only.

Terminology:

Grid cell - is specific, usually square or rectangular, area that is defined by a grid system overlaid on the map. Each cell is identified by its coordinates within the grid.

Species distribution - refers to the geographic location and range of a particular species.

Latitudinal range - refers to the extent of an organism's distribution or the geographic range along the Earth's latitude lines.

Climatic projection - future climate conditions based on mathematical models and simulations.

Bioclimatic variables - are environmental factors related to climate that influence the distribution and behavior of living organisms, including plants and animals.

Costs and benefits: There are no significant costs associated with this project on the assumption that students do not get paid for their time. Probably electric consumption of personal computer to carry out the analyze is trivial. This project could potentially benefit/help scientists and policymakers to understand and address the complex interactions between environmental change and the distribution of plant and animal species. This project could shed light for more effective conservation and sustainable management of biodiversity in a rapidly changing climate.

- Defining your data-mining goals

Data-mining goals: 1. Using current bioclimatic variables and elevation data to train a model to predict globally endangered species occurrences in Europe. **2.** Using trained model and bioclimatic variables (future projections) to predict globally endangered species occurrences in the future in Europe. **3.** Determine how and if the distribution of globally endangered species in Europe have changed. **4**. Visualize the result and create a poster from the main results.

Data-mining success criteria: For species distribution modeling (machine learning model) the model accuracy will depend of chosen metric. It is foreseen that model have to perform better than random chance. In case of AUC-ROC a value generally above 0.7 could be considered acceptable.

Task 3. Data understanding

- Gathering data

Outline data requirements:

Dataset 1 (628 MB, GeoTiff format) Bioclimatic variables for nowadays. The bioclimatic variables represent annual trends (e.g., mean annual temperature, annual precipitation) seasonality (e.g., annual range in temperature and precipitation) and extreme environmental factors (e.g., temperature of the coldest and warmest month).

Dataset 2 (451 MB, GeoTiff format) Bioclimatic variables for the future (plan to use data for 2040-2060)

Dataset 3 (17.1 MB, GeoTiff format) Spatial data of elevation for Europe (extracted).

Dataset 4 (359 MB, CSV format) Globally endangered species occurrence data in Europe (extracted). I plan to use species occurrence data from years 2000-2023.

Verify data availability: Data can be downloaded (verified and downloaded) from following links:

Bioclimatic variables for nowadays: (https://www.worldclim.org/data/bioclim.html)

Bioclimatic variables for the future:
(https://www.worldclim.org/data/cmip6/cmip6_clim2.5m.html)

Elevation: (https://www.worldclim.org/data/worldclim21.html)

Species occurrence data: (https://www.gbif.org/occurrence/search)

Define selection criteria:

For bioclimatic variables nowadays it is foreseen to use all the nineteen bioclimatic variables at resolution of 2.5 minutes. As it is global dataset (covers all continents of the world) the whole dataset has to be downloaded as it's not possible to make selection (e.g., selecting Europe continent) beforehand. Selection must be done separately in GIS software.

For bioclimatic variables for the future, it is foreseen to use all the nineteen bioclimatic variables at resolution of 2.5 minutes. It is planned to use data for 2040-2060. Additionally, three different climate projections will be selected (from out of 14) which themselves are divided into four different shared socioeconomic pathways (ssp). These pathways help researchers and policymakers understand the potential impacts of different societal choices on greenhouse gas emissions and, consequently, on future climate conditions. ssp370 will be used in the current

project. This represents a future scenario with high population growth and assumes fragmented and reactive efforts to address sustainability issues, leading to higher greenhouse gas emissions. As it is global dataset (covers all continents of the world) the whole dataset has to be downloaded as it's not possible to make selection (e.g., selecting Europe continent) beforehand. Selection must be done separately in GIS software.

Spatial data of elevation for Europe. As it is global dataset (covers all continents of the world) the whole dataset has to be downloaded as it's not possible to make selection (e.g., selecting Europe continent) beforehand. Selection must be done separately in GIS software.

For species occurrence data, it is possible to define selection criteria before downloading the data. For criteria it is planned to use human observation as basis of every record, only species occurrences within Europe, coordinates (longitude and latitude) of the record which must be present with every record, use only endangered species based on IUCN Global Red List Category, occurrence status is present and year range is between start of 2000 and end of 2023.

- Describing data

For bioclimatic variables nowadays all the nineteen bioclimatic variables have same dimensions for raster: 8640 x 4320 (columns x rows) with pixel type as floating point. Coordinate system for datasets is GCS_WGS_1984. Actual values are for terrestrial land only, marine areas have "nodata" values. This data is suitable for current project.

For bioclimatic variables for the future (all the nineteen bioclimatic variables) and elevation data (all have same source) everything above also applies for these datasets. Therefore, these datasets are suitable for current project.

Species occurrence dataset has 663073 rows and 50 columns (csv file). From this only four fields are actually needed for the project: "kingdom", "species", "decimalLatitude" and "decimalLongitude". As all these are present in this dataset therefore it is suitable for current project.

- Exploring data

For bioclimatic variables (current and future) only data considering Europe is explored here as this is the focus of the project. Variables are the same for all the bioclimatic variables:

BIO1 = Annual Mean Temperature

BIO2 = Mean Diurnal Range (Mean of monthly (max temp - min temp))

BIO3 = Isothermality (BIO2/BIO7) (×100)

BIO4 = Temperature Seasonality (standard deviation ×100)

BIO5 = Max Temperature of Warmest Month

BIO6 = Min Temperature of Coldest Month

BIO7 = Temperature Annual Range (BIO5-BIO6)

BIO8 = Mean Temperature of Wettest Quarter

BIO9 = Mean Temperature of Driest Quarter

BIO10 = Mean Temperature of Warmest Quarter

BIO11 = Mean Temperature of Coldest Quarter

BIO12 = Annual Precipitation

BIO13 = Precipitation of Wettest Month

BIO14 = Precipitation of Driest Month

BIO15 = Precipitation Seasonality (Coefficient of Variation)

BIO16 = Precipitation of Wettest Quarter

BIO17 = Precipitation of Driest Quarter

BIO18 = Precipitation of Warmest Quarter

BIO19 = Precipitation of Coldest Quarter

Table 1. Basic statistics for bioclimatic variables (both nowadays and future projections) and elevation.

	Bioclimatic variable																			
	BIO1	BIO2	BIO3	BIO4	BIO5	BIO6	BIO7	BIO8	BIO9	BIO10	BIO11	BIO12	BIO13	BIO14	BIO15	BIO16	BIO17	BIO18	BIO19	Elevation
Current																				
Minimum	-12.75	0.00	0.00	0.00	-0.41	-26.36	0.00	-18.75	-12.77	-1.87	-23.28	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-2.84
Maximum	19.13	14.64	47.07	1069.47	35.47	12.29	39.50	21.26	26.74	26.74	15.15	3391.05	414.90	142.00	95.36	1171.67	470.86	591.34	1072.83	2293.57
Mean	6.73	7.83	29.65	689.24	21.09	-5.31	26.39	10.29	4.64	15.39	-1.43	746.27	92.93	36.25	30.11	255.92	122.09	192.48	185.02	302.52
Standard Deviation	5.64	2.10	6.87	186.34	6.71	6.75	6.46	7.16	9.15	5.48	6.50	345.35	41.37	19.51	12.68	117.46	62.13	83.96	117.69	315.19
NULLs	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Future_proj_1																				
Minimum	-7.50	1.40	10.87	253.20	1.58	-19.40	9.50	-12.51	-10.05	0.08	-16.10	228.02	29.00	0.00	9.08	80.93	6.73	6.73	61.78	-2.84
Maximum	20.70	15.46	46.75	1037.00	38.46	12.87	37.56	24.70	29.79	29.00	15.53	3643.65	441.45	164.38	97.63	1202.51	524.64	725.53	1199.94	2293.57
Mean	9.31	8.05	31.10	671.43	23.94	-1.97	25.92	12.34	7.88	17.86	1.68	779.14	100.37	37.78	31.87	271.26	126.55	197.95	198.84	302.52
Standard Deviation	5.23	2.19	6.04	152.94	6.47	5.66	5.33	7.02	9.83	5.26	5.57	348.74	45.14	19.62	12.39	120.00	61.35	84.32	127.61	315.19
NULLs	137.00	137.00	137.00	137.00	137.00	137.00	137.00	137.00	137.00	137.00	137.00	137.00	137.00	137.00	137.00	137.00	137.00	137.00	137.00	0.00
Future_proj_2																				
Minimum	-7.07	1.40	12.80	256.50	-0.40	-17.86	9.70	-12.02	-10.11	-0.29	-14.47	214.50	29.17	0.00	8.86	82.37	6.73	6.73	67.76	-2.84
Maximum	20.60	14.98	48.95	1083.51	38.58	12.40	37.99	23.66	29.90	29.00	15.23	3402.53	410.89	135.25	95.20	1221.73	463.90	600.67	1046.57	2293.57
Mean	9.03	7.87	29.58	692.56	23.66	-2.77	26.75	11.95	7.75	17.81	0.91	766.72	95.56	36.89	30.31	262.26	125.34	191.73	194.44	302.52
Standard Deviation	5.06	2.04	5.29	165.49	6.93	5.52	5.91	6.85	10.01	5.53	5.40	340.39	42.31	18.78	12.22	119.86	60.29	80.69	118.76	315.19
NULLs	137.00	137.00	137.00	137.00	137.00	137.00	137.00	137.00	137.00	137.00	137.00	137.00	137.00	137.00	137.00	137.00	137.00	137.00	137.00	0.00
Future_proj_3																				
Minimum	0.13	-0.74	-7.37	315.70	7.46	-13.76	9.64	-1.25	-7.92	5.86	-7.87	209.82	30.10	0.00	10.22	83.94	6.73	7.00	65.24	-2.84
Maximum	22.20	16.45	48.80	1057.35	40.96	14.47	38.55	26.11	31.40	30.60	17.30	3492.93	436.73	153.65	97.53	1296.65	492.66	645.00	872.39	2293.57
Mean	11.73	8.07	28.81	711.92	27.58	-0.04	27.62	14.54	10.86	20.93	3.55	759.35	95.15	36.83	29.80	261.17	124.66	185.20	188.12	302.52
Standard Deviation	4.22	2.56	7.17	150.92	6.22	4.73	5.78	6.06	10.61	4.67	4.63	337.87	44.94	18.57	12.88	125.80	58.61	79.52	108.20	315.19
NULLs	137.00	137.00	137.00	137.00	137.00	137.00	137.00	137.00	137.00	137.00	137.00	137.00	137.00	137.00	137.00	137.00	137.00	137.00	137.00	0.00

From species occurrence data only four fields are actually needed for the project:

Kingdom - used to categorize living organisms based on their fundamental characteristics. This column has three text values: Animalia, Fungi and Plantae.

Species - assigned a two-part scientific name, comprising a genus name and a species epithet. Altogether there is 655 different species in the dataset.

DecimalLatitude – showing latitudinal value for area where species was found in decimals. Lowest value in this column is 35.01 and highest value is 79.48.

DecimalLongitude - showing longitudinal value for area where species was found in decimals. Lowest value in this column is -31.26 and highest value is 39.66.

- Verifying data quality

There are no major data quality issues. There seems to be small mismatch in current and future bioclimatic variables extent, there is 137 grid cells which have no data for future projections but have data for current bioclimatic variables. Possible solution would be using only the grid cells (in analyzes) where all datasets have data for in sense of bioclimatic variables.

As species occurrence data is from the range 2000-2023, it will need some work, as originally, same species can be present in the same grid cell multiple times due to recording in different years. For the current project it is important to record species presence only once, either present or not during that timeframe.

Task 4. Planning your project

As I am only one doing this project, then all the tasks will be made by me.

List of tasks in the project:

1. Create a custom grid cell system for Europe in order to merge information about species occurrences and bioclimatic data (about 1h, done using ArcGIS)
2. Find mean values for every grid cell both for current/future bioclimatic variables and elevation (about 7-10h, done using ArcGIS)
3. Recalculate/locate species occurrences based on custom grid cell system (about 3-4h, done using ArcGIS).
4. Create an output for every species in order to use these as inputs for modelling (time is going to depend based on how many species will be used finally in analyzes, therefore around 6-8h, done using ArcGIS)
5. Create a code and run the modelling (predict species occurrences in the feature). (about 5-8h, done using Python with some (not decided yet) machine learning algorithms)

6. Analyze the results and make conclusions by creating some tables/plots/maps (about 5-8h, done using ArcGIS, Python).
7. Prepare the poster for submission (about 5h).
8. Finalize everything (double-checking if everything done as asked) (about 2h)