FORGENIUS

Improving access to FORest GENetic resources Information and services for end-Users

Deliverable D4.5

Multivariate description of genomic, environmental and functional status for a subset of the selected GCUs/species

Planned delivery date (as in DoA):M24 31/12/2022

Actual submission date: M25 13/01/2023

Workpackage: WP4

Workpackage leader: CNR

Deliverable leader: UH

Version: 1.0

Project funded by the European Commission within the Horizon 2020 Programme	
Dissemination Level	
PU Public	**PU**
CI Classified, as referred to Commission Decision 2001/844/EC	
CO Confidential, only for members of the consortium (including the Commission Services)	

Research and Innovation action: GA no. 862221

Start date of the project: January 1st, 2021

TABLE OF CONTENTS

1 Summary

Multivariate descriptors were used to pool together data from genetic, environmental and physiological levels of 21 GCUs. There are strong covariance patterns that need to be taken into account by removing redundant variables and subjecting them to dimension-reduction methods. Analysis shows that correlative patterns and multidimensional indexes can be deduced by suggested methods from the data pooled across WP1, WP2 and WP4. More elaborate analysis will be conducted using data from a larger number of GCUs and species, the production of which is in progress.

2 Introduction

In WP4, the objective is to collect genetic data from a subset of GCUs and combine it with environmental and functional status data to achieve a holistic view on the scale and distribution of variation among GCUs. The purpose is also to identify highly correlated variables and other covariance structure across the GCUs and descriptive variables, which may help to extrapolate knowledge also to less well studied GCUs. The ultimate purpose is to be able to assess GCUs adaptability and resilience. In Task 4.3 different statistical approaches from simple multivariate analysis to more advanced machine learning based methods will be used and validated by both empirical data and simulations to identify most robust and informative way to describe the GCUs. In this deliverable, we present multivariate description of genomic, environmental and functional status for a subset of the 21 GCUs.

3 Results

These analyses are based on a subset of 21 GCUs and 12 species (Table 1) where genetic diversity, environmental and physiological data was available. Genetic diversity and physiological data originate from the H2020 project GenTree (Opgenoorth et al., 2021). Soil data, including coarse fragments, percentage of clay, silt and sand, were extracted from the SoilGrids (Poggio et al., 2021). Daily climate data were derived from ERA5 land data (Muñoz-Sabater et al., 2021) and a 40-year (1981-2020) average of each climatic variable was used.

Table 1 21 GCUs used for the multivariate analysis

Species	GCUCode	Country
Betula pendula	ESP00337	Spain
Fagus sylvatica	FRA00029	France
Picea abies	FRA00092	France
Pinus pinaster	ESP00129	Spain
Pinus pinaster	ESP00186	Spain
Pinus sylvestris	ESP00170	Spain
Pinus sylvestris	FRA00101	France
Pinus sylvestris	GBR00001	United Kingdom
Populus nigra	DEU00140	Germany
Populus nigra	ESP00395	Spain
Populus nigra	ITA00045	Italy

Quercus petraea	ESP00387	Spain
Abies alba	ESP00339	Spain
Abies alba	FRA00004	France
Abies alba	FRA00019	France
Pinus halepensis	ESP00057	Spain
Pinus halepensis	ESP00091	Spain
Pinus halepensis	ESP00377	Spain
Pinus halepensis	ITA00076	Italy
Taxus baccata	ESP00340	Spain
Taxus baccata	provisional10	United Kingdom

3.1 Correlative patterns

Correlative patterns among genetic parameters as well as with environmental and functional data from WP1-2 were explored to identify and describe the key axes of the multivariate space. There were 21 variables, including one genetic variable, 7 phenotypic variables and 13 environmental variables. Multicollinearity is a common phenomenon for a subset of two categories (phenotypic and environmental properties) (Fig. 1). Absolute correlation coefficients greater than 0.95 indicate variables which can be typically considered unacceptable collinearity (O'Brien, 2017). Including the nearly-redundant variables can also cause the principal component analysis (PCA) to overemphasize their contribution. For pairs of variables with correlation coefficients higher than 0.95, the one that had a lager variance inflation factor (VIF) was dropped out. Hence, RHair_max, RHair_min, RHair_mean, Tair_max, Tair_min, p07_canopy_1, p09_crown_ellipse and p10_crown_round (for abbreviations, see Annexes) were excluded. Fig. 2 shows the correlation matrix between the 13 selected variables.

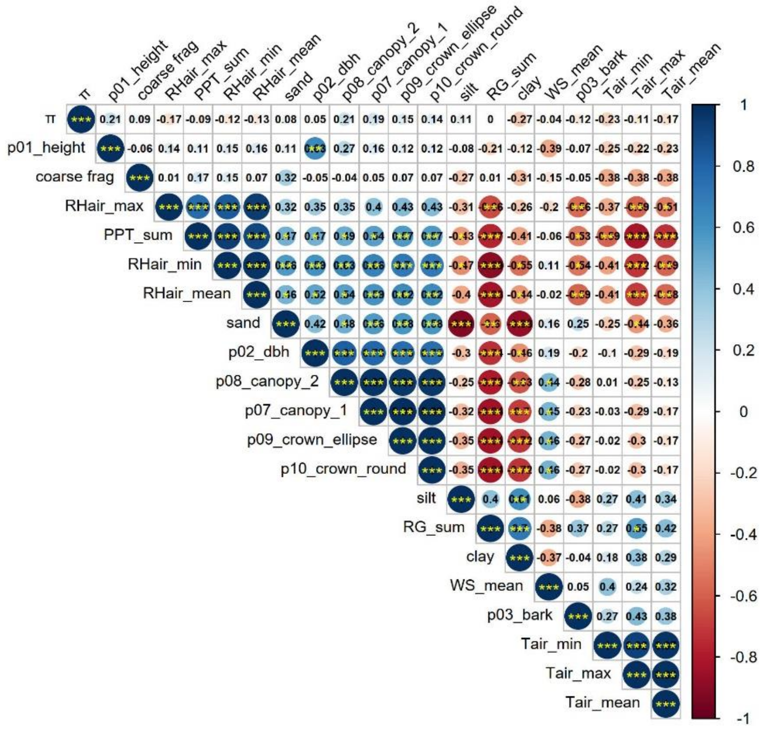

Fig .1 Correlation matrix between 21 variables.

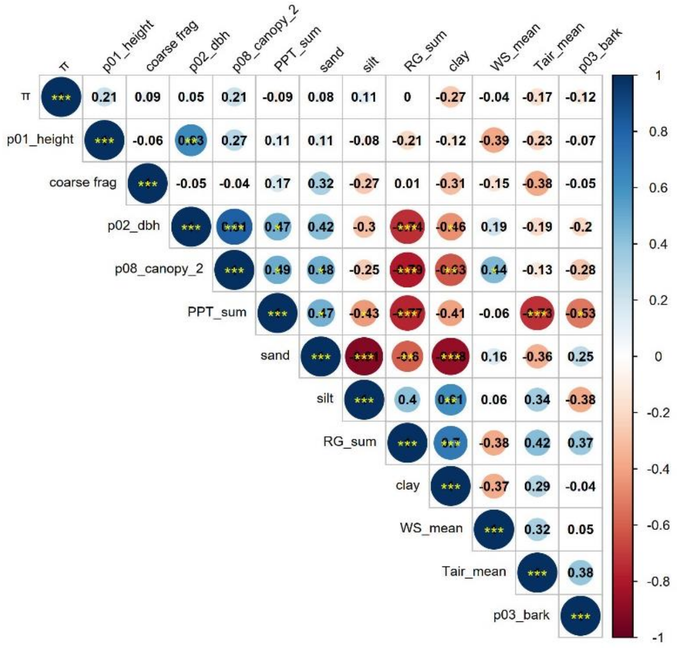

Fig. 2 Correlation matrix between 13 selected variables.

3.2 Principal component analysis

Estimators of genetic diversity, functional and environmental characteristics were submitted to principal component analysis as a single vector of variables per GCU to investigate covariation and develop further composite variables taking into account the correlations among variables.

The first two PCs together (Fig. 3) explains about 54% of the variation. The first three principal components explain 69% of the variation. This is an acceptably large percentage.

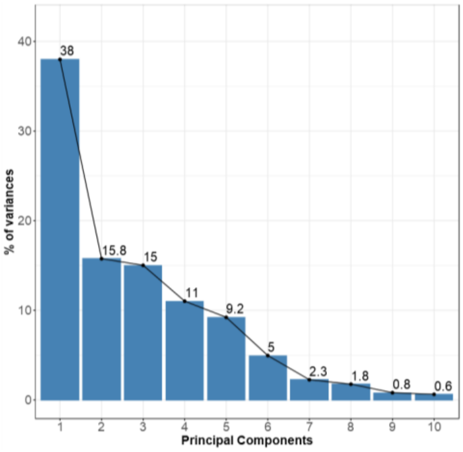

Fig. 3 Amount of the variation (%) explained by each dimension/ principal component (PC).

Fig. 4 shows the contribution of variables to PC1 and PC2. The larger the value of the contribution, the more the variable contributes to the component. RG_sum (i.e. global radiation) had the largest contribution to PC1, followed by the canopy, and by dbh. It also shows the relationships between all variables. Positively correlated variables are grouped, while negatively correlated variables are positioned on opposite sides of the plot origin. For example, the RG_sum is highly negatively correlated with the canopy, and with dbh.

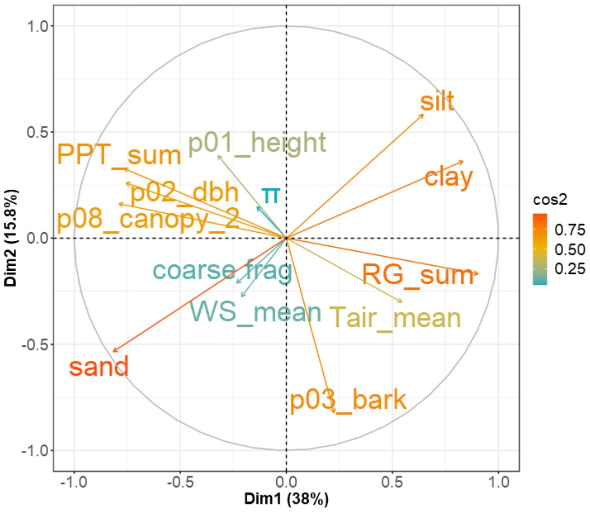

Fig. 4 Contribution of variables to PC1 (Dim 1) and PC2 (Dim 2). A high cos2 indicates a good representation of the variable on the first two PCs.

The contribution to the PCs can also be seen from the position of the variable (Fig. 4). The variable with large contribution is positioned close to the circumference of the correlation circle. The variable with low contribution is close to the centre of the circle. For instance, the nucleotide diversity (π) is very close to the centre of the circle, suggesting that it is less important to interpret the first two PCs. More than two components might be required to better represent the data. This could imply the necessity of more genetic variables for explaining the interaction between genomic, environmental, and functional status.

From Fig. 5, the GCUs closer to each other have similar profiles, whereas those far from each other are dissimilar. The most striking feature is that the two GCUs in the UK were distinguished from the GCUs from other countries at the first principal component. This may underpin the difference between the UK and Continental Europe.

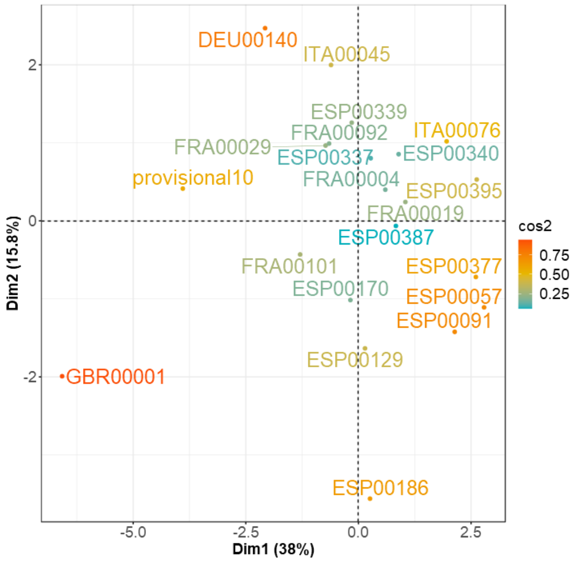

Fig. 5 PCA score plots of the first PC (Dim 1) versus the second PC (Dim 2) of the 21 GCUs. A high cos2 indicates a large contribution of the GCU to the first two PCs.

PCA biplot overlays the score plot (Fig. 4) and the loadings plot (Fig. 5) in a single graph (Fig. 6). A GCU that is on the same side of a given variable has a high value for this variable; a GCU that is on the opposite side of a given variable has a low value for this variable. The value of sand is higher at the GCU GBR00001 than at the other GCUs, which cloud explain why the GBR00001 is far away from the others.

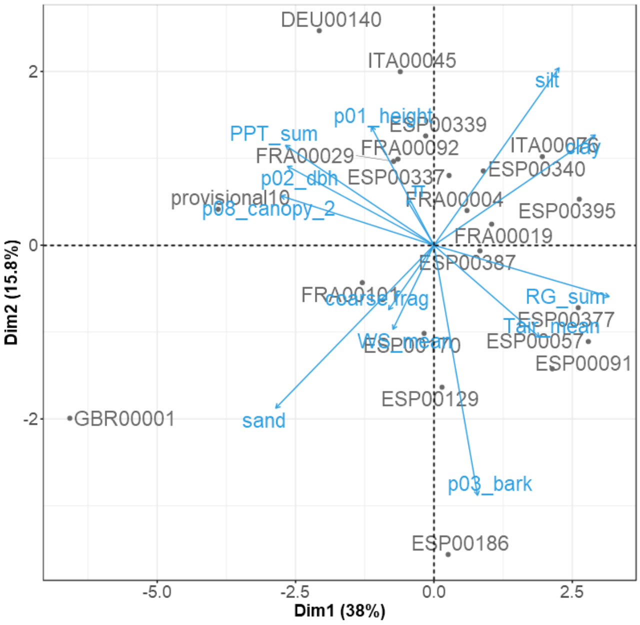

Fig. 6 PCA biplot

3.3 K-means clustering

K-means clustering was used to cluster GCUs in unsupervised manner. Optimal value of 4 clusters were identified by Elbow method (Bholowalia and Kumar, 2014). In Figure 7, it is visible that GCUs do not cluster e.g. by species, but rather by the region or environment, which is not surprising given the high impact of environmental variables evident already e.g. in Figure 4. The GCUs with similar ecological conditions tend to be clustered. For example, the species in cluster 1 generally grow in a cool and somewhat moist environment, while the species in cluster 2 (*Pinus pinaster* only) grow either in Mediterranean or Atlantic climate and the species in cluster 3 generally like a bit warmer climate (except *Abies alba* at FRA00019 which most often grows in cooler climates but there are populations that like warmer climate for this species).

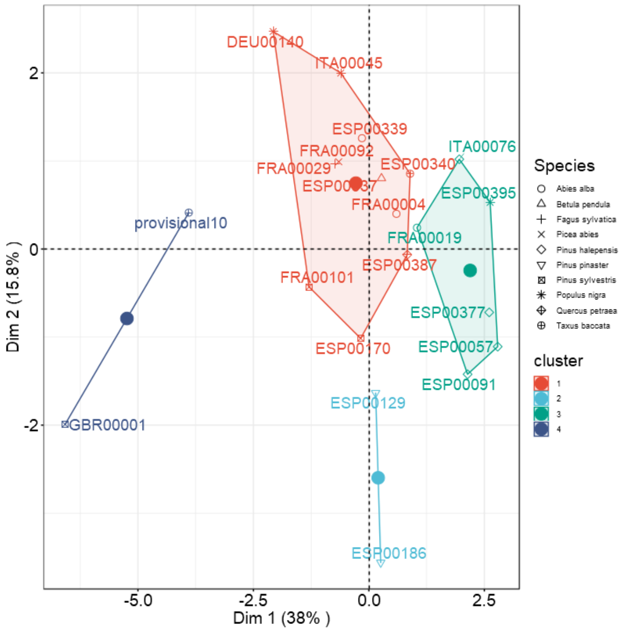

Fig. 7 Results of a K-means Clustering. The full coloured dots represent all GCUs in different clusters.

Table 2 Summary of clusters of the 21 GCUs

cluster	Species	GCUCode	Country
1	*Betula pendula*	ESP00337	Spain
1	*Fagus sylvatica*	FRA00029	France
1	*Picea abies*	FRA00092	France
1	*Pinus sylvestris*	ESP00170	Spain
1	*Pinus sylvestris*	FRA00101	France
1	*Populus nigra*	DEU00140	Germany
1	*Populus nigra*	ITA00045	Italy
1	*Quercus petraea*	ESP00387	Spain
1	*Abies alba*	ESP00339	Spain
1	*Abies alba*	FRA00004	France
1	*Taxus baccata*	ESP00340	Spain
2	*Pinus pinaster*	ESP00129	Spain
2	*Pinus pinaster*	ESP00186	Spain
3	*Populus nigra*	ESP00395	Spain
3	*Abies alba*	FRA00019	France
3	*Pinus halepensis*	ESP00057	Spain
3	*Pinus halepensis*	ESP00091	Spain
3	*Pinus halepensis*	ESP00377	Spain
3	*Pinus halepensis*	ITA00076	Italy
4	*Pinus sylvestris*	GBR00001	United Kingdom
4	*Taxus baccata*	provisional10	United Kingdom

4 Conclusions

In conclusion, multivariate analyses are essential to identify dependencies among different type of variables and our analysis demonstrates that e.g., principal component analysis is a very useful descriptive tool to visualize dependencies. There are strong correlation structures among the variables that need to be taken into account in the future analysis. For example, many environmental variables are nearly redundant and require e.g. dimension reduction before incorporation into indices. These analyses will be repeated later in the project when more shared variable data is available from wider selection of target populations, species and GCUs. Experimental data from WP3 common garden experiments and WP4 simulation studies will be important for validation of relevance of multivariate indices before final index selection.

5 Partners involved in the work

UH, INRAE, CNR, CREAF, GIS

6 Annexes

List of abbreviations.

π	nucleotide diversity
Tair_min	daily minimum air temperature
Tair_max	daily maximum air temperature
Tair_mean	daily mean air temperature
RG_sum	daily global radiation
PPT_sum	daily mean precipitation
RHair_min	daily minimum relative humidity
RHair_max	daily maximum relative humidity
RHair_mean	daily mean relative humidity
WS_mean	wind speed
coarse frag	coarse fragments
p01_height	tree height, m
p02_dbh	diameter at breast height
p03_bark	mean value of bark thickness
p07_canopy_1	canopy projection REP 1
p08_canopy_2	canopy projection REP 2
p09_crown_ellipse	crown ellipse
p10_crown_round	crown size

7 References

Bholowalia, P. and Kumar, A., 2014. EBK-means: A clustering technique based on elbow method and k-means in WSN. International Journal of Computer Applications, 105(9): 17-24.

Muñoz-Sabater, J. et al., 2021. ERA5-Land: a state-of-the-art global reanalysis dataset for land applications. Earth Syst. Sci. Data, 13(9): 4349-4383.

O'Brien, R.M., 2017. Dropping Highly Collinear Variables from a Model: Why it Typically is Not a Good Idea*. Social Science Quarterly, 98(1): 360-375.

Opgenoorth, L. et al., 2021. The GenTree Platform: growth traits and tree-level environmental data in 12 European forest tree species. GigaScience, 10(3).

Poggio, L. et al., 2021. SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty. SOIL, 7(1): 217-240.