Electronic Supplementary Material

distantia: an open-source toolset to quantify dissimilarity between multivariate ecological time-series

Blas M. Benito & H. John B. Birks

Contents

1 Examples

1.1 Comparison of pollen sequences dated to the Last Interglacial

The data used in this example comprises 9 publicly available (European Pollen Database, Pangaea, Neotoma) pollen sequences attributed to the Last Interglacial period (Marine Isotope Stage 5e), also called *Eemian*, with a total of 428 samples of the 22 most abundant pollen types. The data are available under the name *eemian.RData* in the *data* folder. It comprises the dataframes *eemian.sites* (see table and map below), describing some properties of the sites (coordinates, the source database, and the original number of samples and pollen types), and *eemian.pollen*, with the pollen counts for each site and sample. Table 1 shows a summary of the datasets, and Table 2 shows several samples of the different sites.

Table 1: Pollen sequences publicly available representing the Last Interglacial in Central Europe.

| Site	Lat. (º)	Lon. (º)	Elev. (m)	Pollen types	Samples	Reference*
Achenhang	47.75	12.19	834	47	19	Grüger (1983)
Glowczyn_G2	52.42	20.21	93	69	36	Niklewski (1968)
Grobern94	51.68	12.45	105	55	33	Kühl et al. (2007)
Jammertal	48.10	9.73	600	58	34	Müller (2000)
Kletnia_Stara	51.63	21.68	139	53	20	Janczyk-Kopikowa (1989)
Krumbach_I	48.23	10.37	547	81	45	Frenzel (1991)
Naklo	53.15	17.60	88	88	106	Noryśkiewicz (1978)
Ostrow	51.47	19.55	225	41	18	Klatkowa and Winter (1990)
W_KASP	52.23	20.98	105	59	65	Krupiński and Morawski (1993)
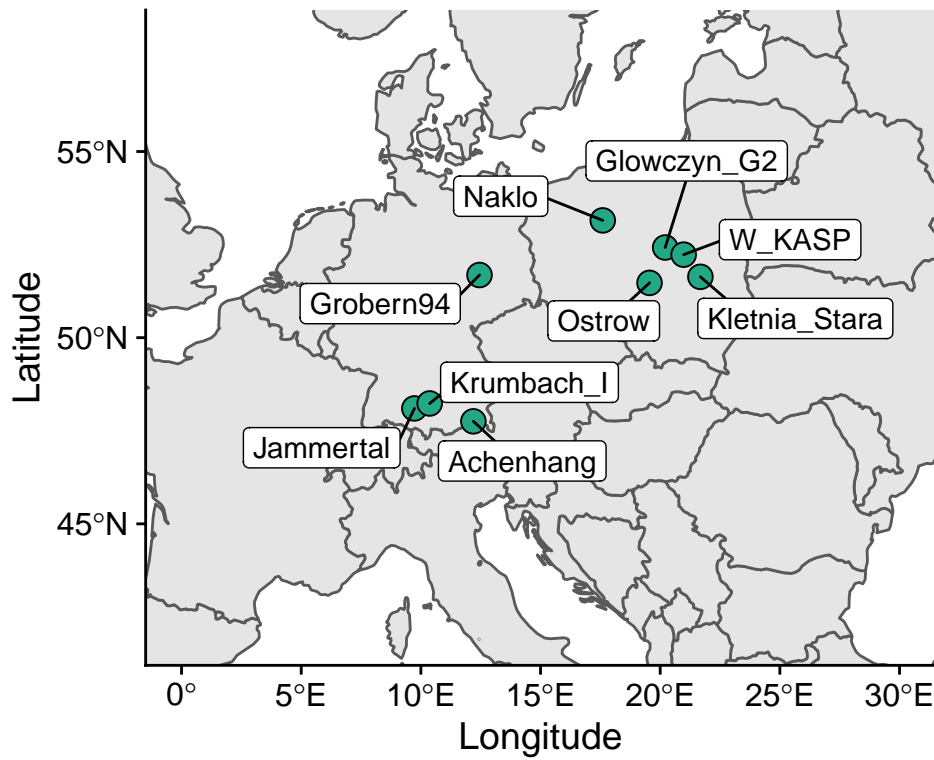
Figure 1: Map of pollen sites used in this case study.

1.1.1 Data preparation

First, we use *prepareSequences()* to apply a Hellinger transformation to the pollen counts, which helps to balance the relative importance of abundant and non-abundant pollen types.

```
ep <- prepareSequences(
  sequences = eemian.pollen,
  grouping.column = "site",
  time.column = "depth",
  transformation = "hellinger",
  paired.samples = FALSE
)
```

1.1.2 Computation of dissimilarity among sequences

Second, we use *workflowPsi()* to compute the dissimilarity between each one of the 36 combinations of sites, and *workflowNullPsi()* to compute the probability of obtaining a given ψ on restricted permutations of the input sequences (null model of the given ψ). The restricted permutation is applied locally, and independently by column. On each data-point, the algorithm decides randomly to either leave it as is, or switch it with the data-point above or below it. The outcome allows to assess the probability of obtaining the observed ψ value by chance on the non-permutated sequences. The result of both functions is shown in *Table 2*.

```
#computing ψ
ep.psi <- workflowPsi(
  sequences = ep,
  diagonal = TRUE,
  grouping.column = "site",
  time.column = "depth",
  method = "euclidean",
  paired.samples = FALSE,
  ignore.blocks = TRUE
)

#computing null psi values (this will take some time!)
ep.psi.null <- workflowNullPsi(
  sequences = ep,
  diagonal = TRUE,
```

```
  grouping.column = "site",
  time.column = "depth",
  method = "euclidean",
  paired.samples = FALSE,
  ignore.blocks = TRUE,
  repetitions = 999,
  parallel.execution = TRUE
)
```

Table 2: Dissimilarity (Psi) between Central European pollen sequences dated to the Last Interglacial.
Column p represents the probability of finding each given Psi by chance.

| A	B	Psi	p
Kletnia_Stara	Ostrow	1.456	0.580
Glowczyn_G2	W_KASP	1.456	0.767
Glowczyn_G2	Kletnia_Stara	1.541	0.857
Jammertal	Krumbach_I	1.574	0.521
Naklo	Ostrow	1.600	0.270
Grobern94	Kletnia_Stara	1.604	0.936
Glowczyn_G2	Naklo	1.671	0.440
Kletnia_Stara	Naklo	1.692	0.141
Grobern94	Naklo	1.722	0.558
Grobern94	Ostrow	1.733	0.615
Kletnia_Stara	W_KASP	1.744	0.819
Krumbach_I	Ostrow	1.805	0.260
Glowczyn_G2	Ostrow	1.839	0.602
Ostrow	W_KASP	1.853	0.364
Glowczyn_G2	Grobern94	1.863	0.993
Achenhang	Ostrow	1.888	0.776
Jammertal	Ostrow	1.907	0.923
Naklo	W_KASP	1.918	0.583
Glowczyn_G2	Jammertal	1.979	0.995
Grobern94	Jammertal	1.989	1.000
Jammertal	Kletnia_Stara	1.995	0.991
Glowczyn_G2	Krumbach_I	2.069	0.635
Jammertal	Naklo	2.074	0.291
Kletnia_Stara	Krumbach_I	2.106	0.801
Jammertal	W_KASP	2.116	0.894
Achenhang	Kletnia_Stara	2.125	0.977
Grobern94	W_KASP	2.154	1.000
Grobern94	Krumbach_I	2.266	0.775
Krumbach_I	W_KASP	2.354	0.705
Krumbach_I	Naklo	2.411	0.598
Achenhang	Krumbach_I	2.431	0.455
Achenhang	Jammertal	2.507	0.986
Achenhang	Grobern94	2.611	0.972
Achenhang	W_KASP	2.752	0.975
Achenhang	Glowczyn_G2	2.814	0.959
Achenhang	Naklo	2.829	0.426
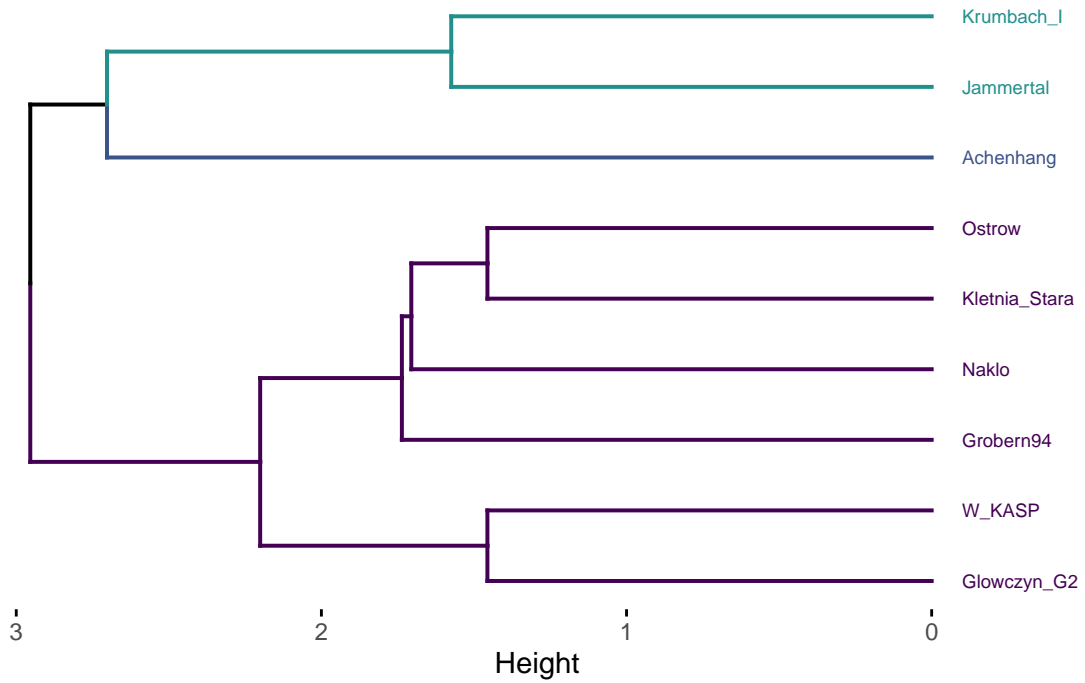
Figure 2: Clustering of the Eemian pollen sites based on the dissimilarity matrix.

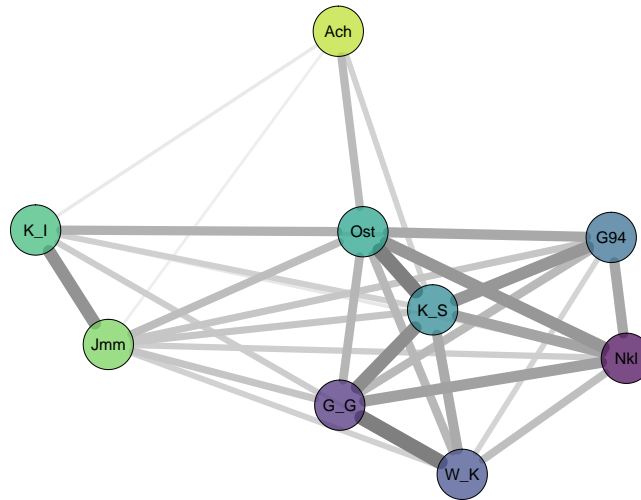Figure 3: Similarity between pollen sites. Colors represent latitude, and edge thickness represent similarity, computed as 1/dissimilarity to highlight stronger similarities.

Figure 6 shows that there is a large cluster of sites that share similarity with Klentia Stara (K_S), a closely related cluster composed by Krumbach I (K_I) and Jammertal (Jmm), and an isolated group, Achenhang (Ach), which is highly dissimilar to the other sites, probably due to the presence of two hiatuses in the sequence (Grüger 1983).

1.1.3 Relationship between dissimilarity and geographic distance between sites.

We hypothesize that pollen assemblages during the Eemian follow the main axiom of geography, and therefore, sites that are close to each other should have higher similarity than sites that are far apart from each other. To evaluate this hypothesis, we first have to compute the geographic distances between sites from their coordinates and differences in elevation. Second, we model dissimilarity as a function of distance, difference in elevation, and difference in latitude with a GLM with Gaussian errors. Model selection is based on AIC values of the model considering all variables and the models with each variable.

Table 3: Standardized coefficients of the GLM model.

Variable	Units	Coefficient	CI.95	p-value
(Intercept)		-0.00	[-0.22;0.22]	1.0000000
distance		-0.03	[-0.40;0.34]	0.8653647
elev.diff		1.00	[0.48;1.53]	0.0006753
lat.diff		-0.25	[-0.87;0.38]	0.4477745

Elevation difference has the higher standardized residual, and its confidence interval does not overlap with zero. Below we fit one model per variable, and compare AICs of all fitted models.

Table 4: AIC values of the complete model (first row) and each one of the univariate models.

	df	AIC
All variables	5	78.97
Distance	3	98.68
Elevation difference	3	76.22
Latitude difference	3	88.52

According to the AIC results, the model based on the elevation difference is the one better explaining ψ. The plot below shows the model fit and the data.
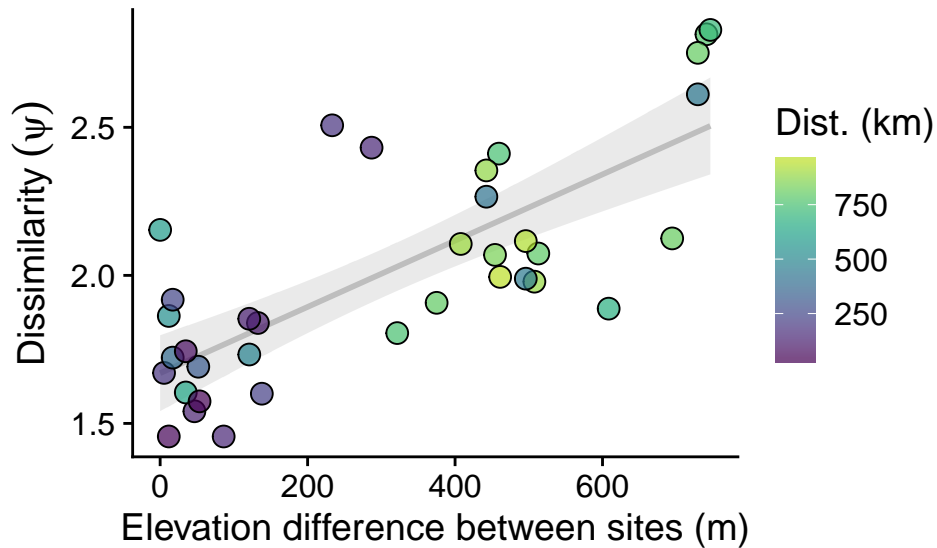
Figure 4: Relationship between dissimilarity and difference in elevation between pollen sites.

1.1.4 Comparing two sites

The sites Jammertal and Achenhang, are among the sites with a more disproportionate relationship between dissimilarity ($\psi = 2.507$, p = 0.986), distance (188 kilometres) and elevation difference (233 m). Considering their geographical positions, why are these sites that different? The function *workflowImportance* allows the assessment of the drop in dissimilarity (in percentage of the dissimilarity when all variables are considered) when a given variable is removed from the data. Variables with a positive psi-drop contribute positively to the dissimilarity between two sequences (the compared sequences become more similar when the variable is removed). The code below shows how the function is applied to both sequences.

```r
#computing importance of each variable
ep.importance <- workflowImportance(
  sequences = ep[ep$site %in% c(A, B), ],
  grouping.column = "site",
  time.column = "depth",
  method = "euclidean",
  paired.samples = FALSE,
  diagonal = TRUE,
  ignore.blocks = TRUE
)
```

Table 5: Drop in dissimilarity (expressed as a percentage of the dissimilarity when computed with all variables) when omitting each variable of the sequences Jammertal and Achenhang

	Psi drop (%)
Picea	22.77
Corylus	9.91
Carpinus	7.00
Abies	5.76
Frangula	0.00
Ranunculaceae	-0.08
Betula	-0.16
Chenopodiaceae	-0.25
Tilia	-0.27
Salix	-0.29
Apiaceae	-0.39
Artemisia	-0.43
Asteraceae	-0.44
Alnus	-0.56
Juniperus	-0.61
Pinus	-0.73
Ulmus	-0.84
Fraxinus	-1.24
Cyperaceae	-1.45
Poaceae	-1.72
Quercus	-4.66
Taxus	-5.65

According to the table, the following pollen types are the ones contributing the most to the dissimilarity between Jammertal and Achenhang: Picea, Corylus, Carpinus, Abies.
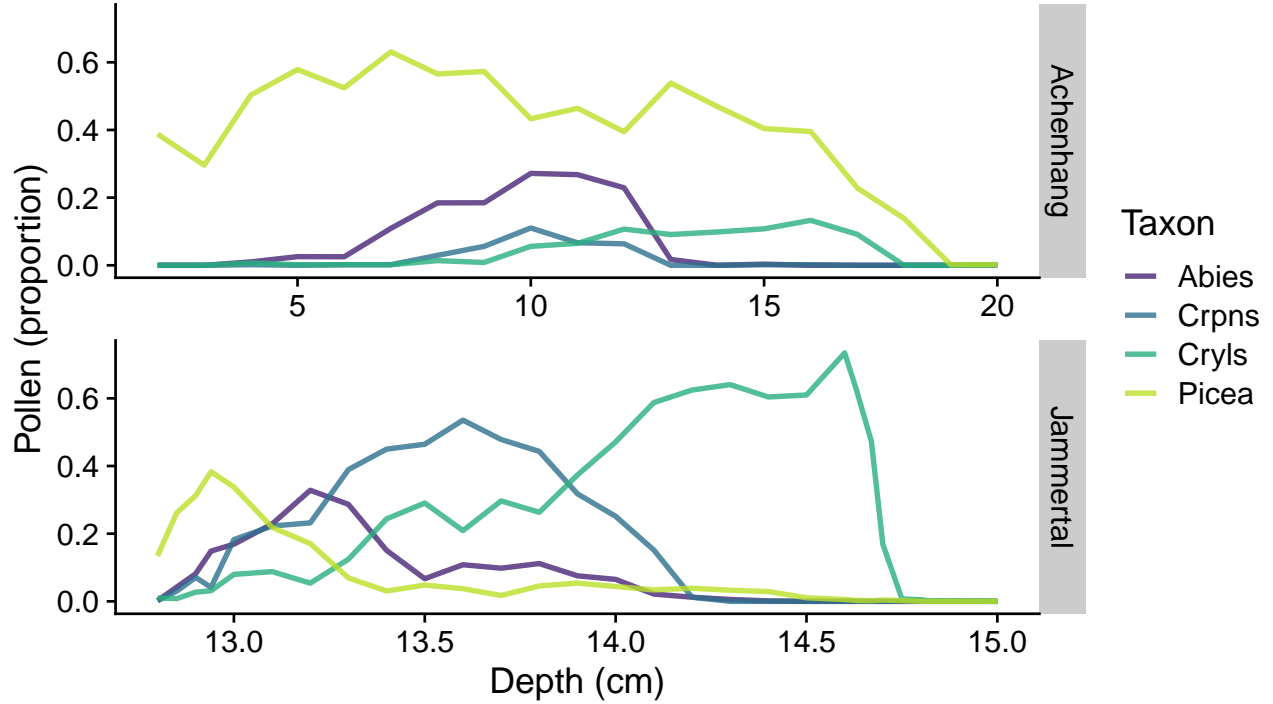
Figure 5: Pollen types (Abies, Carpinus, Corylus, and Picea) contributing the most to the dissimilarity between Achenhang and Jammertal. The samples of these sequences are not aligned in depth, and therefore potential shape distortions may apply when comparing them.

1.2 Functional dynamics of three populations of Fagus sylvatica

1.2.1 Data

For this example we selected three sites that, according the EU Forest database (Mauri et al. 2017), are mono-specific stands of *Fagus sylvatica*. Two are located very close to the coast, one in Northern Spain and another in Southern Sweden, while the remaining one has a continental setting in Western Germany.

Table 6: Coordinates and elevation of three Fagus sylvatica stands in Spain, Germany, and Sweden.

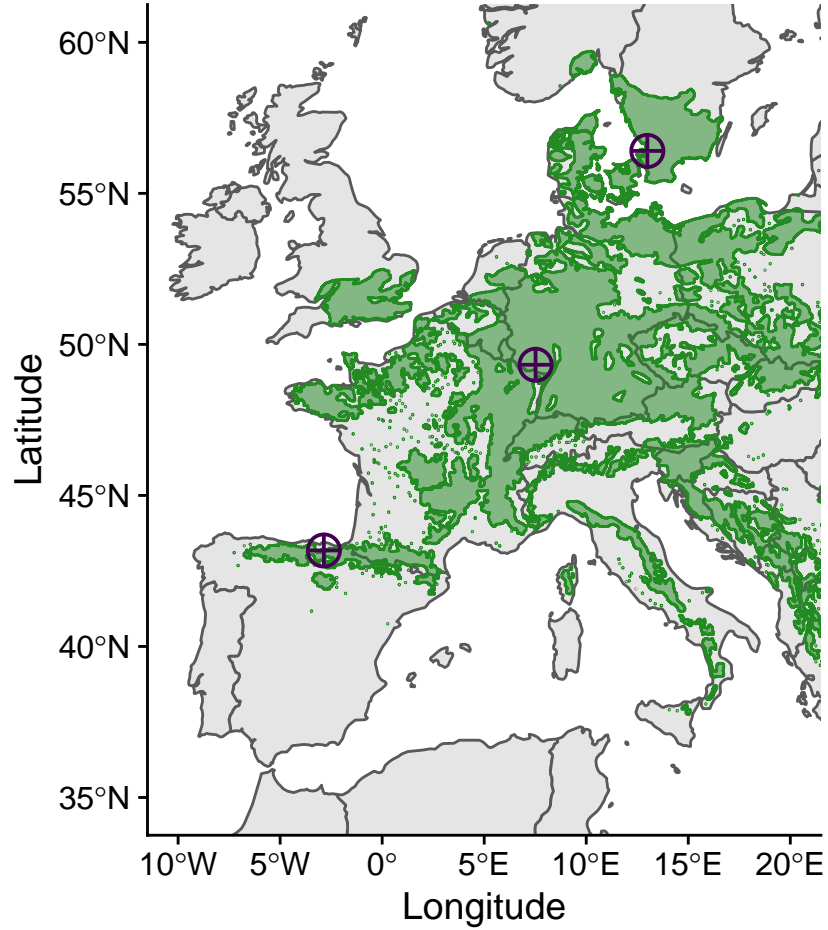| Site	Longitude (º)	Latitude (º)	Elevation (m)
Spain	-2.857	43.177	305
Germany	7.517	49.325	334
Sweden	12.995	56.405	176

Figure 6: Locations of three selected stands of Fagus sylvatica in Sweden, Germany, and Spain. Locations were extracted from the EU-Forest database (Mauri et al. 2017). Green polygons represent the distribution of Fagus sylvatica in Europe, according to EUFORGEN (von Wuehlisch, 2008)

We used the MODISTools R package (Tuck et al. 2014) to retrieve EVI (enhanced vegetation index) values for these sites between the years 2001 and 2018 captured by the MODIS platform. The data comprise two values per month, and to aggregate to months we took the maximum EVI value recorded each month. We then coupled the data with monthly temperature and rainfall data from CRU (Harris et al. 2014).
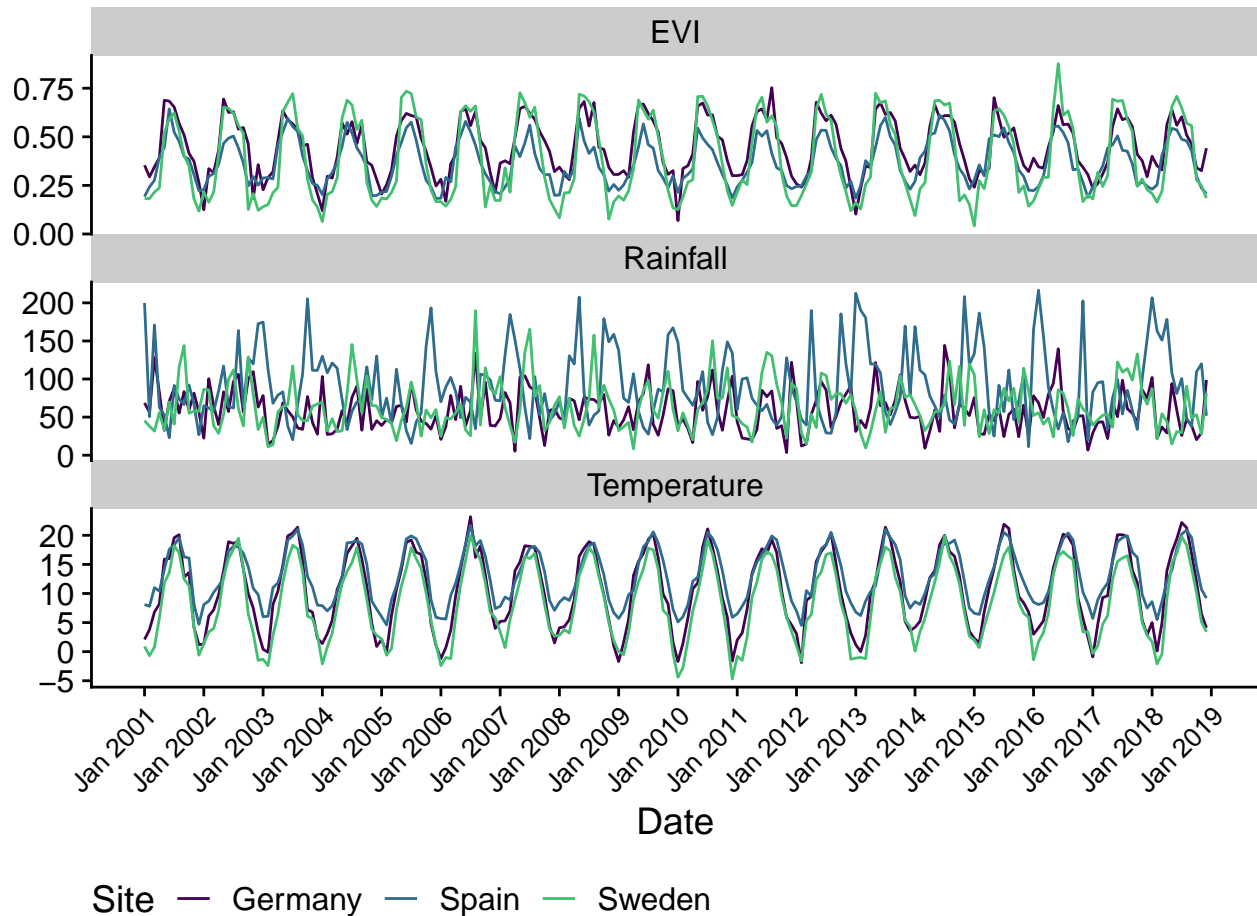
Figure 7: Monthly EVI, rainfall, and temperature of three Fagus sylvatica forests in different countries.

1.2.2 Comparison of complete sequences

The general objective is to understand how the dynamics of the system varies among sites. A first approach consists of comparing the three complete sequences. Below we apply a comparison by shape using the *workflowPsi()* function, and compute ψ values on restricted permutations of the datasets to assess to what extent the dissimilarity between sequences can be due to chance.

```
#scaling data
ec <- prepareSequences(
  sequences = evi.climate,
  grouping.column = "site",
  exclude.columns = c("year", "month", "date"),
  transformation = "scale"
)
```

```
#compute psi
ec.psi <- workflowPsi(
  sequences = ec,
  grouping.column = "site",
  exclude.columns = c("year", "month", "date"),
  method = "euclidean",
  paired.samples = FALSE,
  diagonal = TRUE,
  ignore.blocks = TRUE
)

#psi on restricted permutations
ec.psi.null <- workflowNullPsi(
  sequences = ec,
  grouping.column = "site",
  exclude.columns = c("year", "month", "date"),
  method = "euclidean",
  paired.samples = FALSE,
  diagonal = TRUE,
  ignore.blocks = TRUE,
  repetitions = 999
)
```

Table 7: Dissimilarity in temperature, rainfall, and EVI, between Fagus sylvatica populations in Spain, Germany, and Sweden. Psi represents dissimilarity, Psi NULL is the null expectation of psi, computed as the average of 999 psi values obtained from restricted permutations of the input datasets, and p represents the proportion the psi-null values that were lower than the psi obtained on the observed datasets.

Sites		Psi	Psi-null	p
Spain	Germany	1.2845	1.3502	0.075
Spain	Sweden	1.4009	1.4112	0.607
Germany	Sweden	0.8317	1.0060	0.001

According to the ψ values, the sites from Germany and Sweden have the highest similarity, followed by Spain and Germany. The comparison between Spain and Sweden yielded a ψ value above the

null expectation (average of ψ values computed on restricted permutations of the data), and therefore the results of this comparison can be considered unreliable. It is important to take in mind that the restricted permutation applied by *workflowNullPsi()* only moves randomly selected datapoints one cell up or down, independently by column, and therefore it may yield relatively high p values and low null expectations even when the signal is relatively strong. Therefore, p values and null expectations should be considered as general indicatives of the strength of the relationship, and not as a representation of statistical significance.

1.2.3 Contribution of each variable to dissimilarity

Below we use the function *workflowImportance()* to assess how the variables temperature, rainfall, and EVI contribute to the dissimilarity between the dynamics of the three populations.

```
#importance of variables
ec.psi.importance <- workflowImportance(
  sequences = ec,
  grouping.column = "site",
  exclude.columns =  c("year", "month", "date", "site.year"),
  method = "euclidean",
  paired.samples = FALSE,
  diagonal = TRUE,
  ignore.blocks = TRUE
)
```

Table 8: Percentage of dissimilarity loss when removing a given variable from the data. The ψ value computed on the complete datasets is taken as reference. Positive numbers indicate that the variable contributes to dissimilarity, while negative numbers indicate that the variable contributes to similarity.

Sites		Temperature	Rainfall	EVI
Spain	Germany	11.46	11.66	12.71
Spain	Sweden	10.18	7.20	10.94
Germany	Sweden	-0.70	3.57	10.66

The table above shows the effect on dissimilarity of removing one variable at a time from the compared datasets. EVI plays an important role in increasing the dissimilarity between the German and the Swedish sites. All variables contribute to the dissimilarity between Spain and Germany. Regarding Spain - Sweden comparison, EVI seems to be the variable contributing most to the similarity between

14

sites, but these have been confirmed to be dissimilar beyond the null expectation, and therefore the output of the importance analysis might not be accurate.

1.2.4 Comparison of dynamics year to year

How does the dissimilarity between the dynamics of each pair of sites change over the years? Were they equally dissimilar each year within the study period? If so, what variables contributed the most to the dissimilarity?

The code below groups the data by site and year, computes ψ for each combination of these, and organizes the output to represent it graphically.

```r
#grouping by site and year
ec$site.year <- paste(
  ec$site,
  ec$year,
  sep = "-"
  )


#computation of psi
ec.psi.year <- workflowPsi(
  sequences = ec,
  grouping.column = "site.year",
  exclude.columns = c("site", "year", "month", "date"),
  method = "euclidean",
  paired.samples = FALSE,
  diagonal = TRUE,
  ignore.blocks = TRUE
)
```
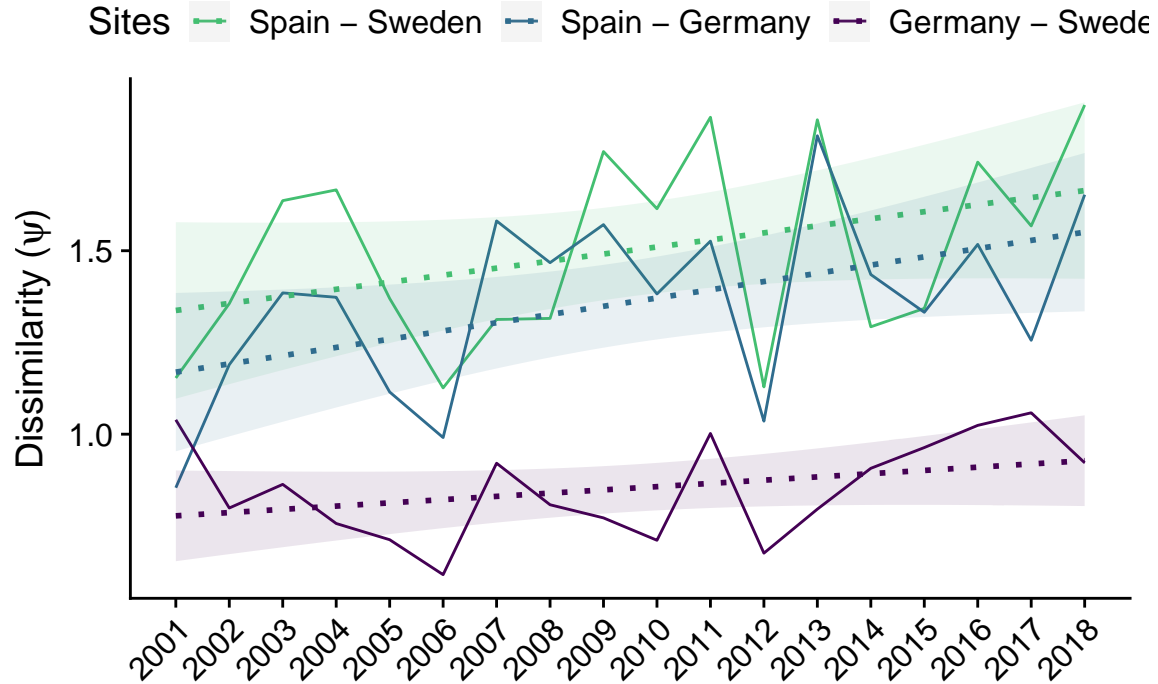
Figure 8: Dissimilarity between Fagus sylvatica stands by year.

According to these results, there are several events that could have reduced the dissimilarity between sites, and specially between the Spanish and German sites. These are the cold and heat waves of 2006, the cold wave of 2012, and the extreme drought in Spain that started in 2013, the European-wide heat wave of 2015, and the cold wave of 2017.

Even thought the limited amount of data available (18 years only) prevents the assessment of the statistical significance of the temporal trend of dissimilarity between sites, and in spite of the punctuated events affecting dissimilarity, there is a pattern of increasing dissimilarity between the Spanish and the other sites, and an almost neutral decrease in the dissimilarity of the German and the Swedish sites. What variables are contributing to these patterns?
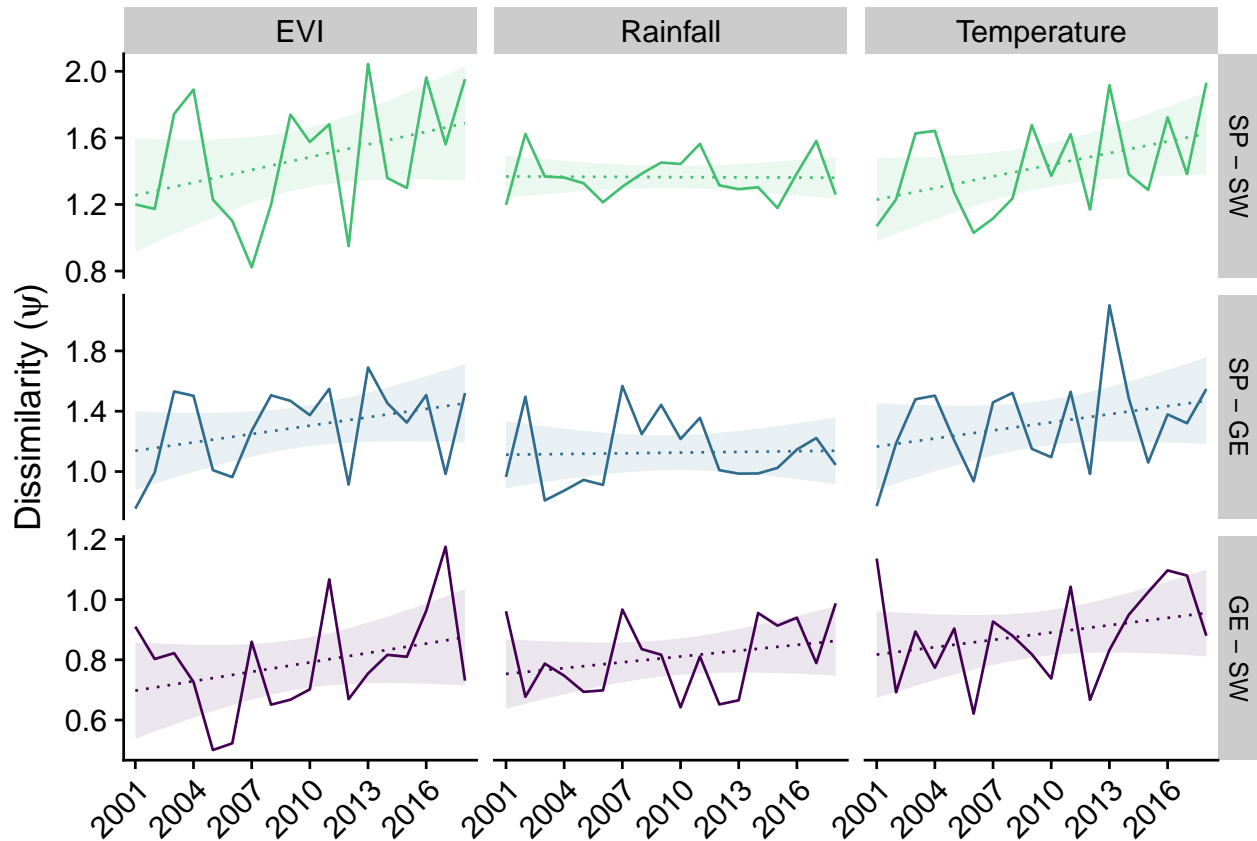
Figure 9: Dissimilarity when removing a given variable. Variables with positive trends are increasing their contribution to similarity over time.

All variables contribute to the increment in dissimilarity between the German and Swedish. In the case of Spain vs. Sweden and Germany, rainfall is the variable with the highest contribution to the increment in similarity between sites, even though the signal is weak.

These dissimilarity trends can be explained by changes in the EVI and temperature values of each month across years. Particularly, the EVI values of the Swedish populations are increasing across most months except the colder ones, while spring and summer EVI values are being lower and lower each year in the Spanish and German populations. Increased summer temperatures in these populations would likely be behind this response, but the data available per month and year are too scarce to allow a robust assessment of significance.
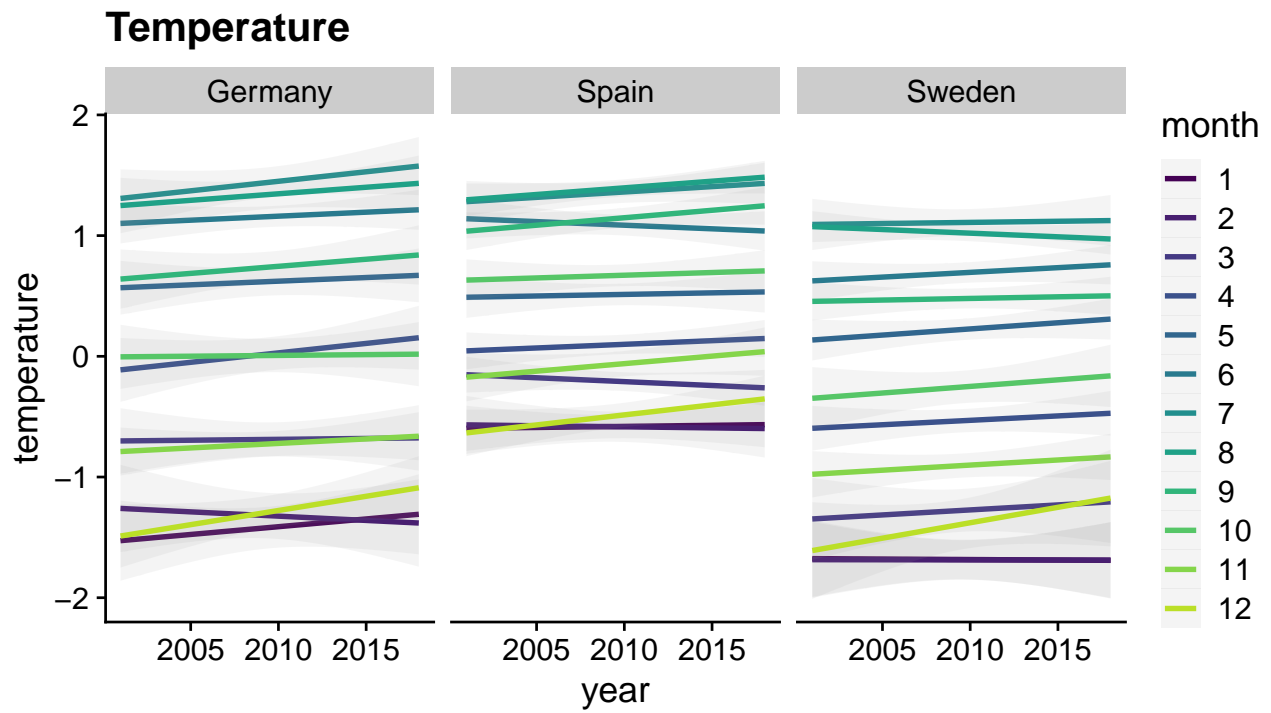
Temperature

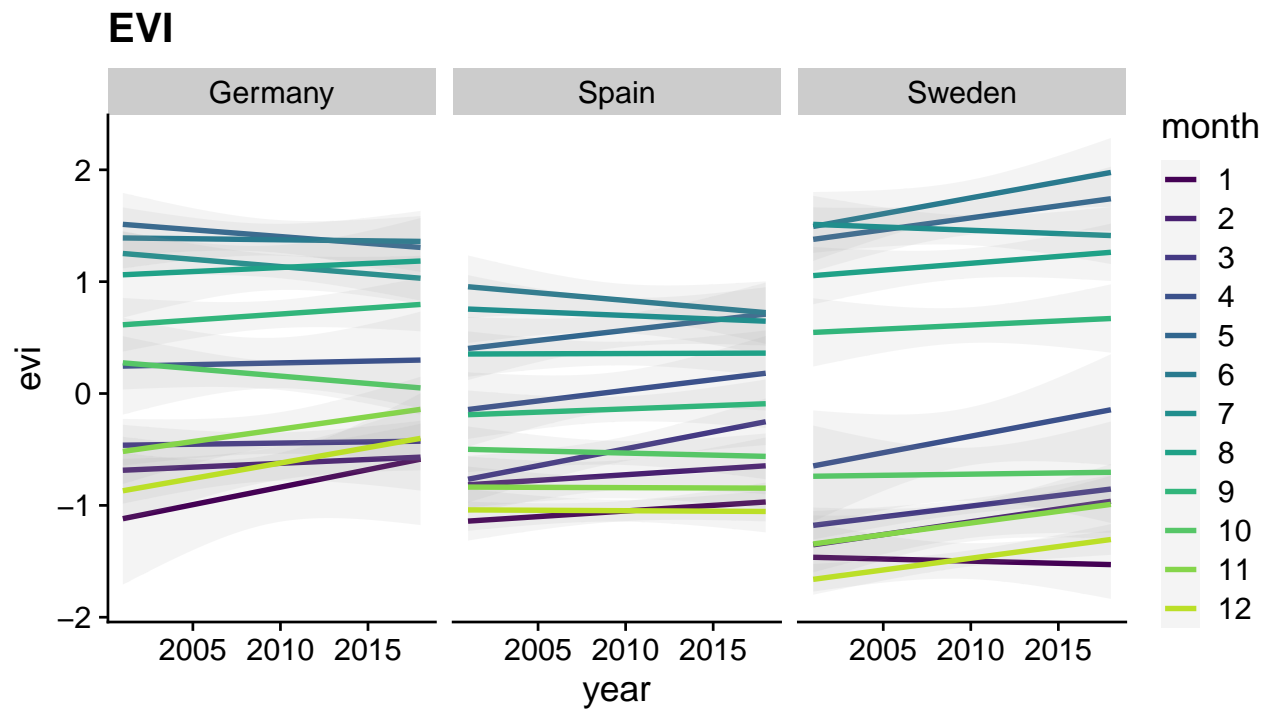Figure 10: Temperature trends by month per site and year.

EVI

Figure 11: EVI trends by month per site and year.
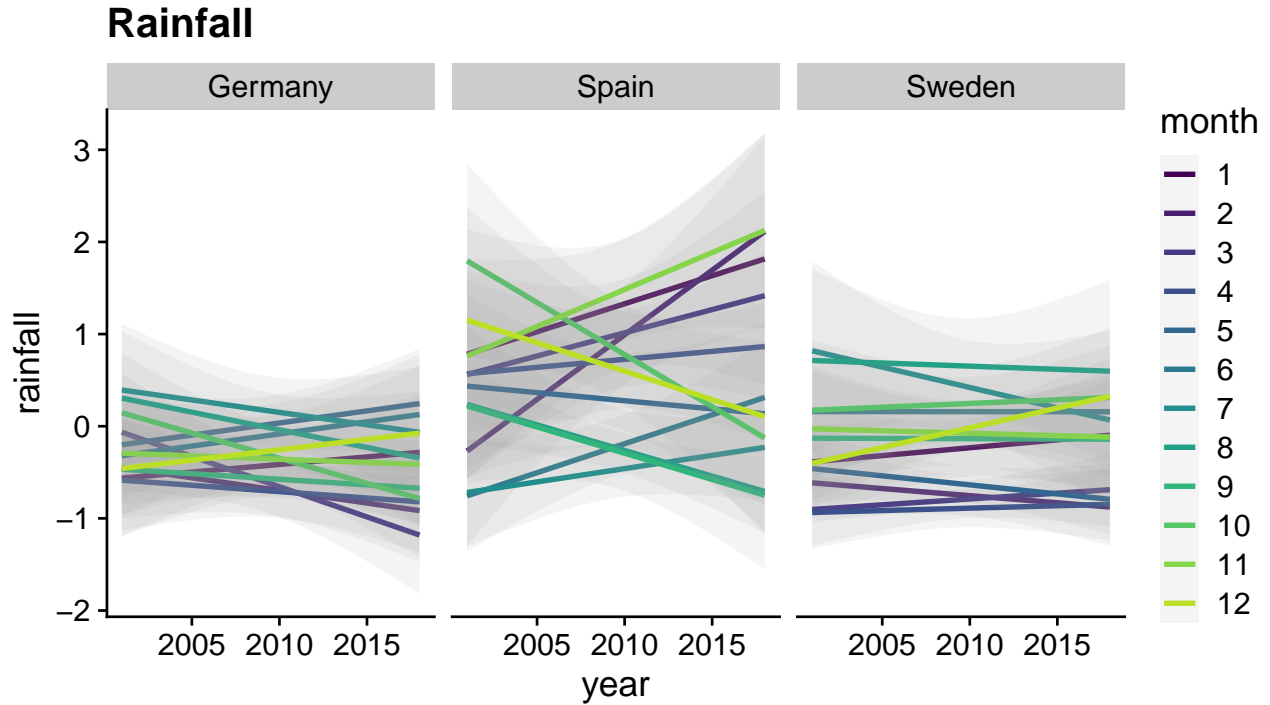
Rainfall

Figure 12: Rainfall trends by month per site and year.

2 Sensitivity of the dissimilarity metrics implemented in *distantia*

The *distantia* package implements an elastic dissimilarity measure with two distance-minimization options: elastic (comparison by shape) with diagonal and non-diagonal search of the least cost path, and a lock-step (comparison by time). It also implements several distance metrics, of which Euclidean and Manhattan are the most general ones. Furthermore, depending on the nature of the data, input sequences can be transformed with one of the following options: "scaled" (scaling and centering), "proportion", "hellinger" (square root of the proportions).

In this section we explore the effect of these different options on the outcome of the dissimilarity analyses that can be carried out with the *distantia* package. The analyses will be based on two example datasets provided with the package: *climate*, and *pollenGP*. The former is a climatic dataset with 800 observations of 4 variables with different units (average temperature and rainfall, and temperatures of the warmest and coldest month) generated by a palaeoclimatic simulation (Willeit et al. 2019). The data transformations applied to this dataset will be "none" (the data with their natural values), and "scaled" (scaling and centering through the *scale()* function of the R software). *PollenGP* is a pollen dataset with 200 observations of 40 pollen types (Woillard, 1979). In this case we use the

data transformations "proportion" (relative abundance of each taxa) and "hellinger" (square root of the proportions).

Two different scenarios of potential differences between the datasets and modified copies of themselves will be explored.

In **Scenario 1** we increase the differences between a dataset and its copy by modifying an increasing number of datapoints (intersection between a row and a column). This analysis involves creating a copy of a dataset, selecting a variable, selecting a case, changing its value, and comparing the original dataset with its copy through different combinations of method (elastic diagonal, elastic non-diagonal, lock-step), transformation ("none" and "scaled" for *climate*, "proportion" and "hellinger" for *pollenGP*), and distance metrics (Manhattan and Euclidean). The process is repeated a 30 times for each number of datapoints modified, and the mean and standard deviation of the resulting ψ values are reported.

In **Scenario 2** we remove an increasing number of randomly selected rows of the copy dataset to assess how differences in sample size impact the dissimilarity between two otherwise identical datasets through combinations of method (lock-step and elastic with and without diagonals), transformation ("none" and "scaled" for *climate*, "proportion" and "hellinger" for *pollenGP*), and distance metrics (Euclidean and Manhattan).

These operations are performed by the functions available in the file *sensitivity_functions.R* The function *repetitionsPsi()* computes ψ on a given dataset and its modified copy a number of repetitions, while *sensitivityPsi()* runs each whole experiment.

The code below prepares the parameters of the experiment through the custom function *generateParams()*, which, as *repetitionsPsi()* and *sensitivityPsi()*, is not part of the *distantia* package, but can be found in the file *sensitivity_functions.R* provided as companion of the .Rmd version of this file.

```r
#loading functions
source("sensitivity_functions.R")


#loading the data
data("climate")
data("pollenGP")
pollen <- pollenGP[ ,1:18]
rm(pollenGP)


#removing unneeded columns
```

```
climate$sequenceId <- NULL

climate$time <- NULL

pollen$depth <- NULL

pollen$age <- NULL


#generating parameters file
#function generateParams is in sensitivity_functions.R
params.climate <- generateParams(
  data = climate,
  transformations = c("none", "scale")
  )


params.pollen <- generateParams(
  data = pollen,
  transformations = c("proportion", "hellinger")
  )
```

The sensitivity analysis below computes ψ 30 times (repetitions) for each combination of parameters contained in the parameters dataframes. On each repetition, different data-points are modified for **Scenario 1**, and different rows are removed for **Scenario 2**. ψ values obtained from the sensitivity analysis are aggregated through repetitions, and the mean and standard deviation across repetitions are available in the output dataframe of the sensitivity analysis.

```
#sensitivity for climate dataset
sensitivity.climate <- sensitivityPsi(
  data = climate,
  params = params.climate,
  repetitions = 30
  )


#sensitivity for pollen dataset
sensitivity.pollen <- sensitivityPsi(
  data = pollen,
  params = params.pollen,
  repetitions = 30
  )
```

21

2.0.1 Scenario 1: effect of number of modified data-points

We plot changes in ψ mean and standard deviation across number of data-points modified for each method and combination of transformation and distance metric.
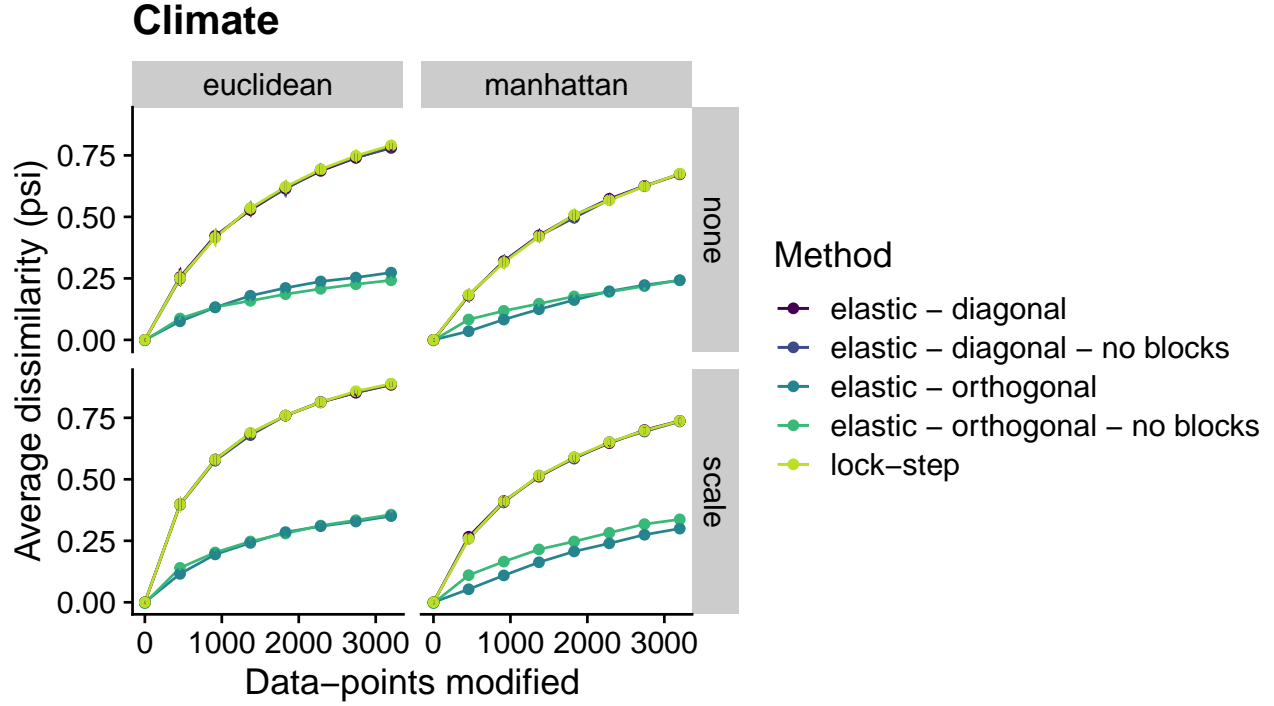

Figure 13: Psi mean and standard deviation across number of data-points modified for the climate dataset.
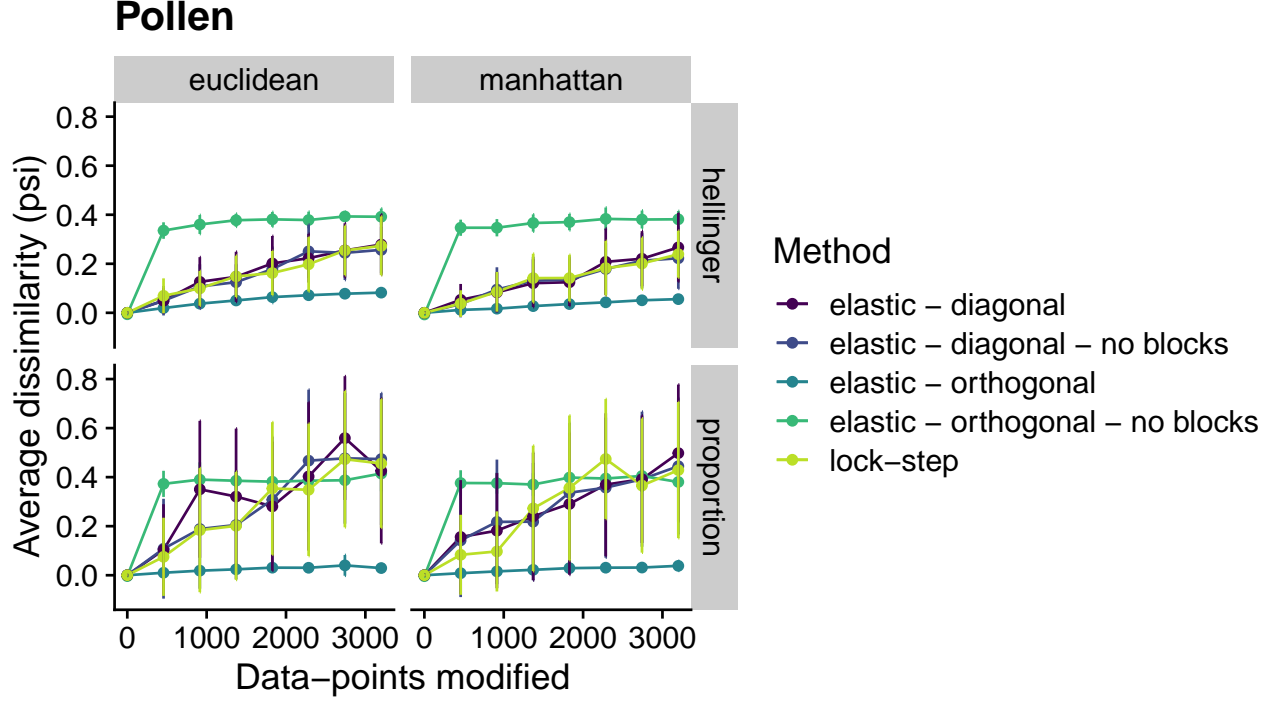
Figure 14: ψ mean and standard deviation across number of data-points modified for the pollen dataset.

We found that lock-step and elastic-diagonal methods to compute dissimilarity are practically equivalent, specially when the data are scaled and the differences between datasets are small. Both methods were, as expected, more sensitive to dissimilarity than the elastic-orthogonal method, especially when differences between the datasets were small, and the distance metric used was Euclidean. The adaptive nature of the elastic-orthogonal method also increased the standard deviation of ψ values with larger differences between the compared datasets.

2.0.2 Scenario 2: effect of difference in sample size between the compared sequences

The analysis of sensitivity of elastic measures to changes in the number of rows of one of the compared datasets shows that ψ increases exponentially with increasing number of samples removed from the copy dataset. This finding indicates that, even when two sequences have the same overall dynamics, their dissimilarity scores can be highly inflated, particularly when one of the sequences has less than half the number of samples of the other sequence. The orthogonal method showed to be less sensitive to this effect, while transformation and distance metric seemed to be irrelevant.
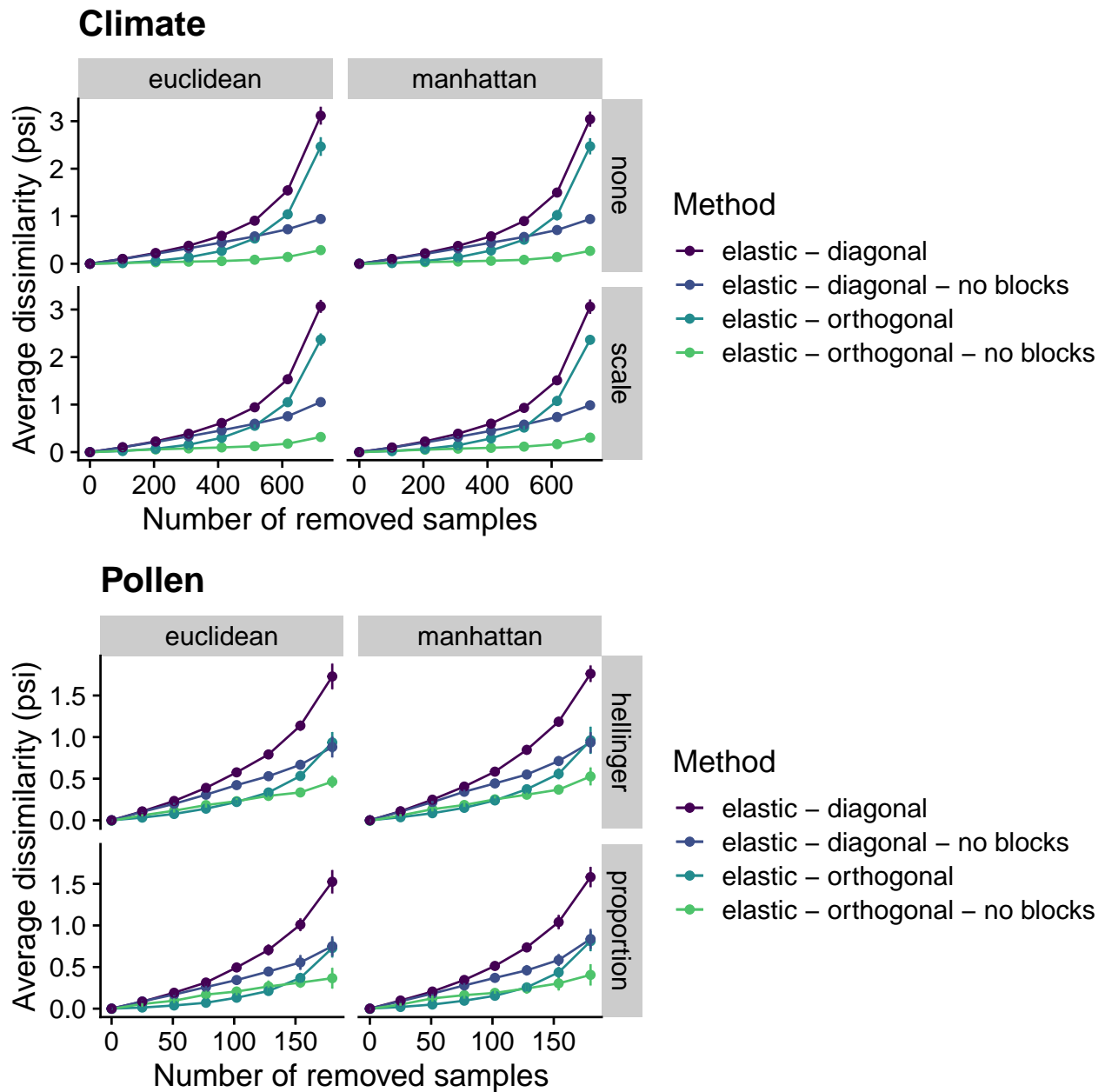
Figure 15: ψ mean and standard deviation across number of removed samples for the pollen dataset.

3 References

Willeit, M., Ganopolski, A., Calov, R., & Brovkin, V. (2019). Mid-Pleistocene transition in glacial cycles explained by declining CO2 and regolith removal. Science Advances, 5(4), eaav7337. doi: 10.1126/sciadv.aav7337

Woillard, G. M. (1978). Grande Pile peat bog: A continuous pollen record for the last 140,000 years.

Quaternary Research, 9(1), 1–21. doi:10.1016/0033-5894(78)90079-0

Mauri, A., Strona, G., & San-Miguel-Ayanz, J. (2017). EU-Forest, a high-resolution tree occurrence dataset for Europe. Scientific Data, 4, 160123. doi:10.1038/sdata.2016.123

Tuck et al. (2014). MODISTools - downloading and processing MODIS remotely sensed data in R Ecology & Evolution, 4(24), 4658 - 4668.

Harris, I., Jones, P. D., Osborn, T. J., & Lister, D. H. (2014). Updated high-resolution grids of monthly climatic observations – the CRU TS3.10 Dataset. International Journal of Climatology, 34(3), 623–642. doi:10.1002/joc.3711

von Wuehlisch, G. (2008). EUFORGEN Technical guidelines for genetic conservation and use for European beech. Biodiversity International. Rome, Italy. URL: http://www.euforgen.org/species/fagus-sylvatica/

Grüger E (1983) Untersuchungen zur Gliederung und Vegetationsgeschichte des Mittelpleistozäns am Samerberg in Oberbayern. Geologica Bavarica 84:21-40

Niklewski J (1968) Interglacial eemski w; Glowczynie kolo Wyszogrodu [The Eemian interglacial at Glowczyn near Wyszgrod, central Poland]. Monographiae Botanicae 27:125-191

Kühl N, Litt T, Schölzel C, Hense A (2007) Eemian and Early Weichselian temperature and precipitation variability in northern Germany. Quaternary Science Reviews 26:3311-3317

Müller UC (2000) A Late-Pleistocene pollen sequence from the Jammertal, south-western Germany with particular reference to location and altitude as factors determining Eemian forest composition. Vegetation History and Archaeobotany 9 (2):125-131. doi:10.1007/bf01300062

Janczyk-Kopikowa Z (1989) A new locality of Eemian interglacial deposits near Deblin. Kwartalnik Geologiczny 33 (2):269-275

Frenzel B (1991) Über einen frühen letzteiszeitlichen Vorstoss des Rheingletschers in das deutsche Alpenvorland. Paläoklimaforschung 1:377-400

Noryśkiewicz B (1978) Interglacjał Eemski w Nakle nad Notecią [The Eemian Interglacial at Nakło on the River Noteć (N Poland)]. Acta Palaeobotanica 19 (1):67-112

Klatkowa H, Winter H (1990) The Eemian Interglacial at Ostrow near Grabica. Acta Geographica Lodziensia 61:59-68

Krupiński KM, Morawski W (1993) Geological position and pollen analysis of Eemian interglacial sediments of Warsaw-Wawrzyszew. Acta Palaeobotanica 33 (1):309-346