Supporting information: Crowd-sourced plant occurrence data reveal macroecological patterns

1 Supplementary Text

1.1 Preprocessing of climate and soil data

We used WorldClim (`https://www.worldclim.org/`; Fick and Hijmans, 2017) including both, monthly mean values and precomputed bioclimatic variables and soil physical and chemical properties derived from Soil Grids (`https://soilgrids.org/`; Hengl et al., 2014) data for interpreting the recovered spatial patterns. Both data sets contain many co-linear variables and are preprocessed via principal component analysis (PCA) here. Note that we do not use nonlinear dimensionality reduction in this case, because it was essential to interpret the recovered axes which is not easily possible in the nonlinear case. The PCA can be understood here as a data preprocessing step to make the the subsequent variable selection and prediction with random forests more efficient and accessible to interpretation. Throughout this section we show the results of the PCA always for both, the climate and the soil data as left and right column figures respectively.

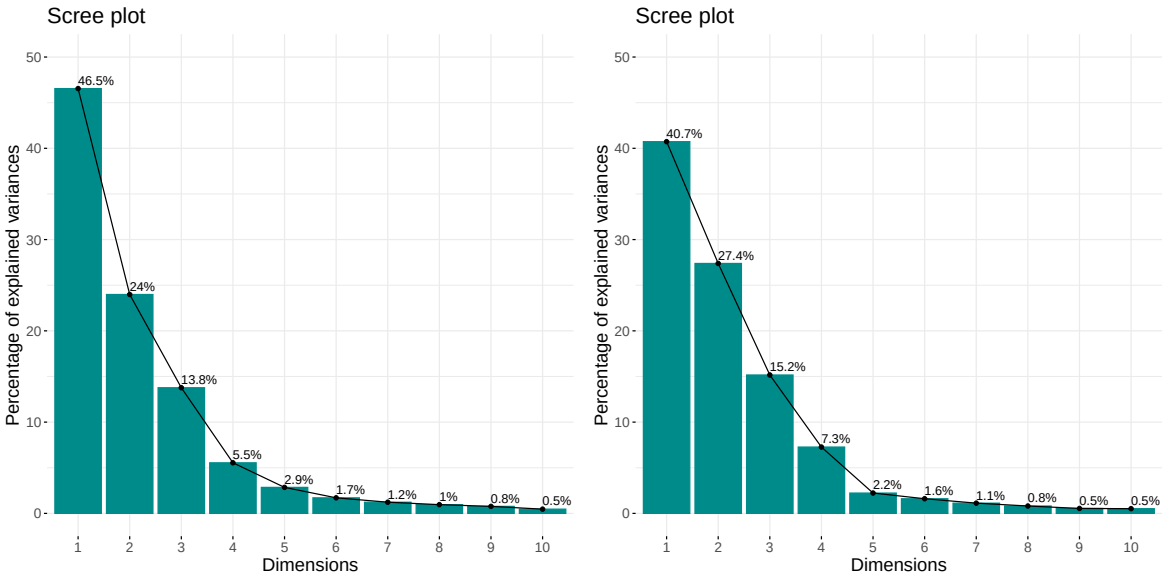

Figure S1: Variances explained by each dimension in for WorldClim (left) and SoilGrids (right) by the leading principal components. After four dimensions the explained variances drop to very small amounts so that we retain five dimensions in both cases.

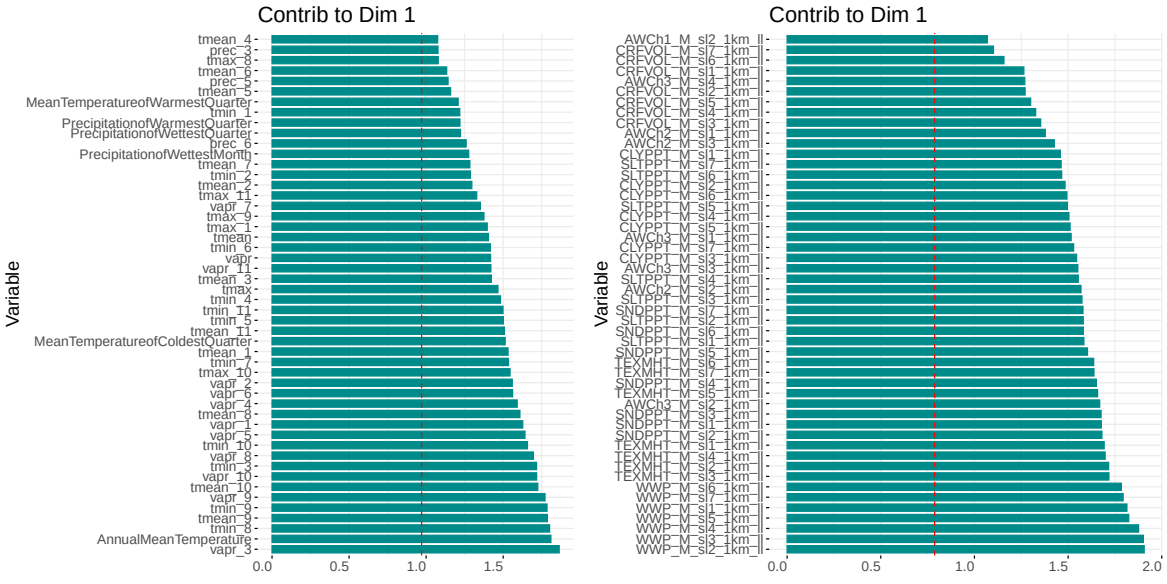

Figure S2: Contributions to the 1^{st} principal component for WorldClim (left) and SoilGrids (right).

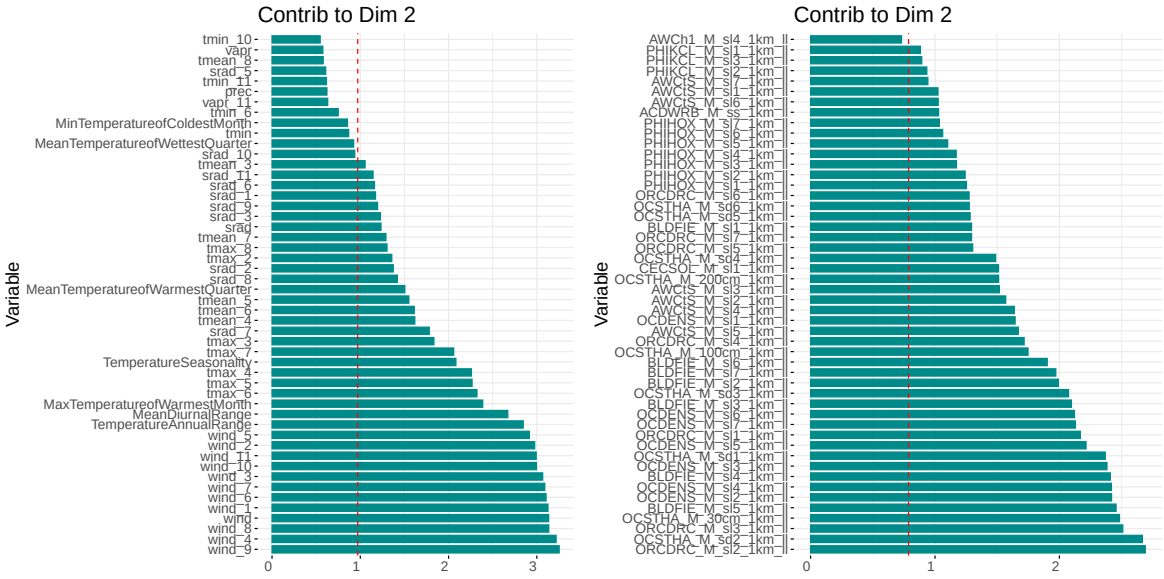

Figure S3: Contributions to the 2^{st} principal component for WorldClim (left) and SoilGrids (right).

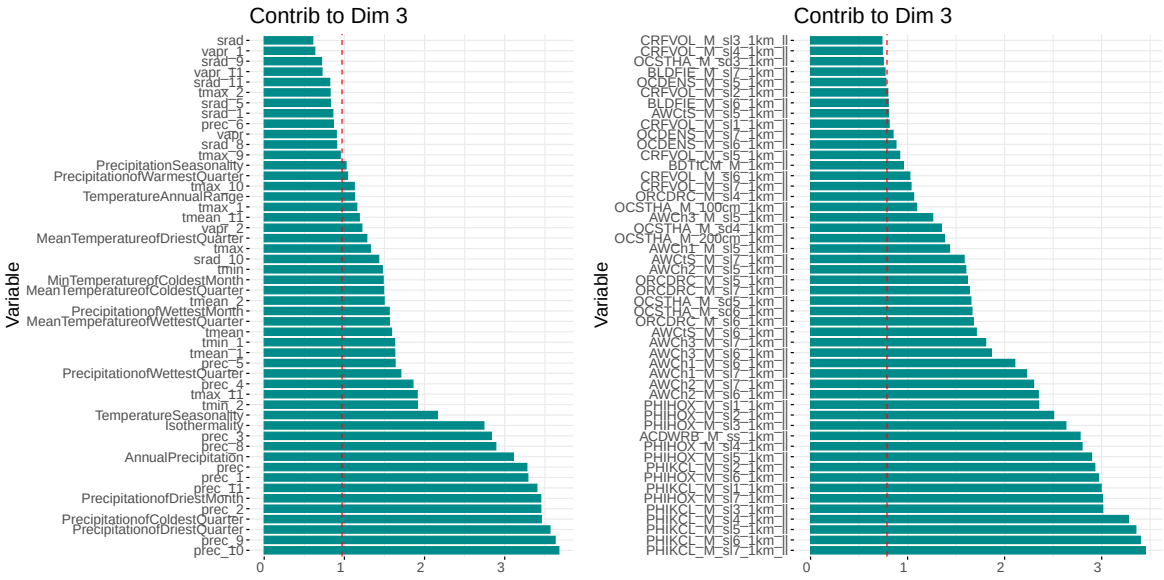

Figure S4: Contributions to the 3^{st} principal component for WorldClim (left) and SoilGrids (right).

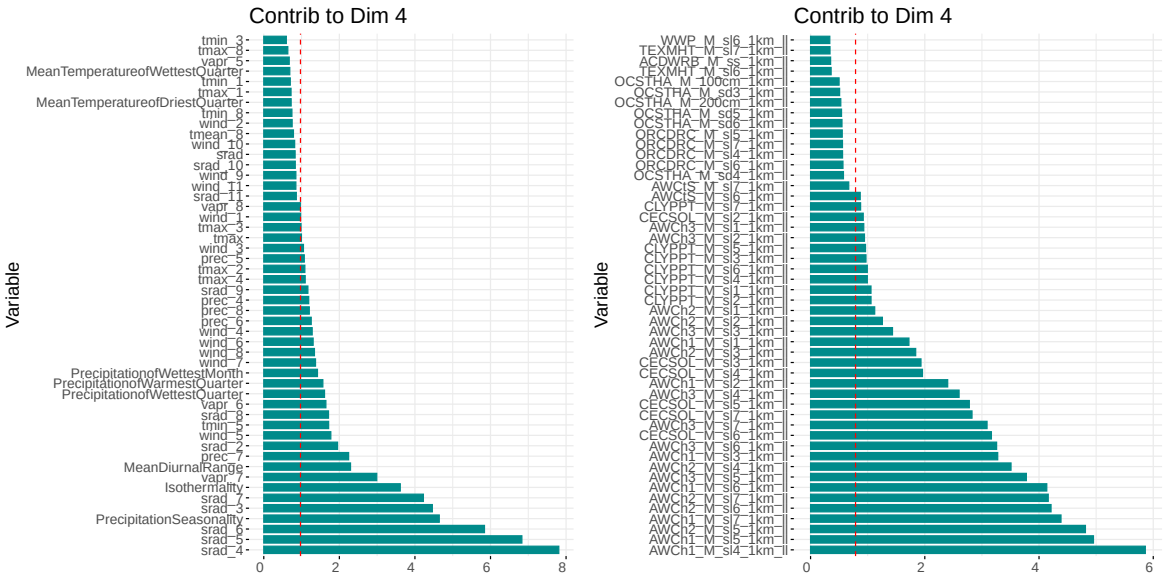

Figure S5: Contributions to the 4^{st} principal component for WorldClim (left) and SoilGrids (right).

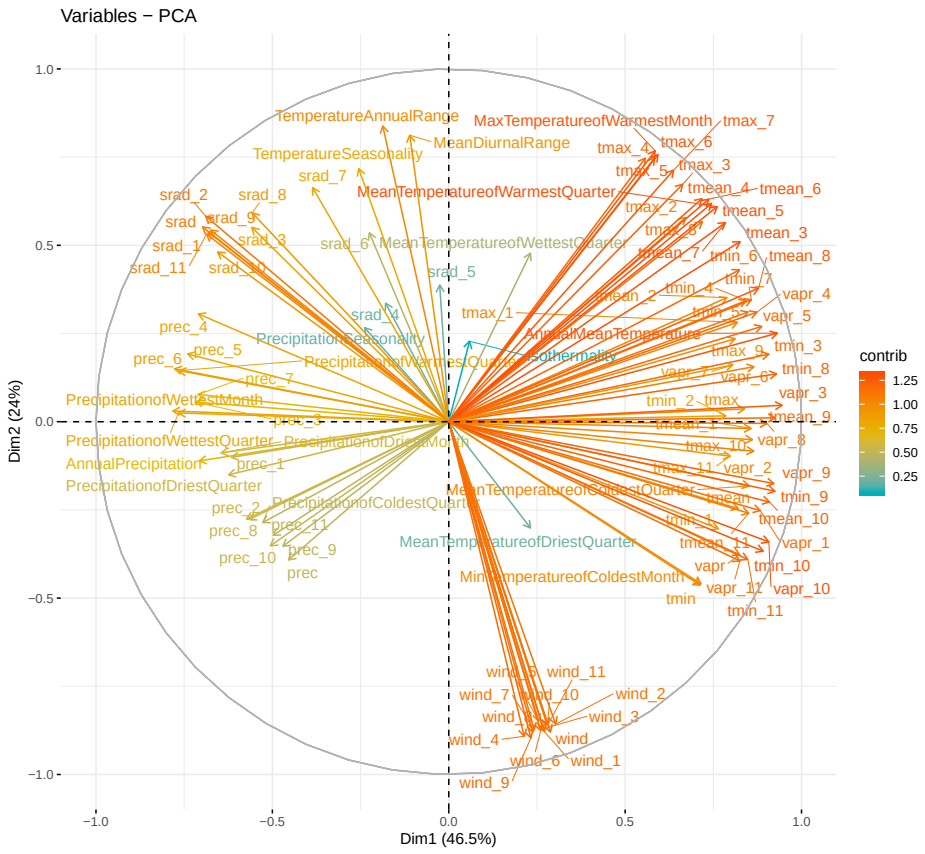

Figure S6: Visualization for the contributions of the considered variables to the first two principal components for WorldClim.

Figure S7: Visualization for the contributions of the considered variables to the first two principal components for SoilGrids.

From inspecting these figures in detail we can infer for WorldClim that over our study area. The aggregation unit were the so-called 'Messtischblätter' (MTB; corresponding to a spatial resolution of 6' $\times$ 10'.The following description holds:

- $\mathrm{PC}_{\mathrm{Clim}}^{(1)}$ = "Temperature axis". Various surface air temperature variables (for instance Annual Mean, monthly values, as well as monthly minima and maxima show up here) but also multiple monthly water vapour pressure (vapr; numbers indicate month) indicators. This makes sense given the intimate relation between water vapour pressure and temperature.

- $\mathrm{PC}_{\mathrm{Clim}}^{(2)}$ = "Wind axis". This axis is controlled by wind speeds off all month. Additionally, it shows high contributions of range variables related to climate amplitudes e.g. Temperatures Annual and Diurnal Ranges, or Seasonality. The refined interpretation would be "Wind and seasonality axis".

- $\mathrm{PC}_{\mathrm{Clim}}^{(3)}$ = "Precipitation axis". Here, we find all monthly precipitation averages as well as the derived bioclimtological variables i.e. Precipitation of Driest Quarter, Precipitation of Coldest Quarter, Precipitation of Driest Month, Annual Precipitation, or Precipitation of Wettest Quarter. Some other variables like Isothermality or Temperature Seasonality are also related, but at the lower edges.

- $\mathrm{PC}_{\mathrm{Clim}}^{(4)}$ = "Radiation axis". Here, relatively high contributions of summer radiation values only.

From inspecting these figures in detail we can infer for SoilGrids that over our study area (Germany at the MTB level aggregation) the following description holds:

- $\mathrm{PC}_{\mathrm{Soil}}^{(1)}$ = "Soil water-texture axis". We find high values of Avaiable Soil Water Capacity (WWP) and textures in terms of sand, silt, clay contents (TEXMHT), as well as sandy

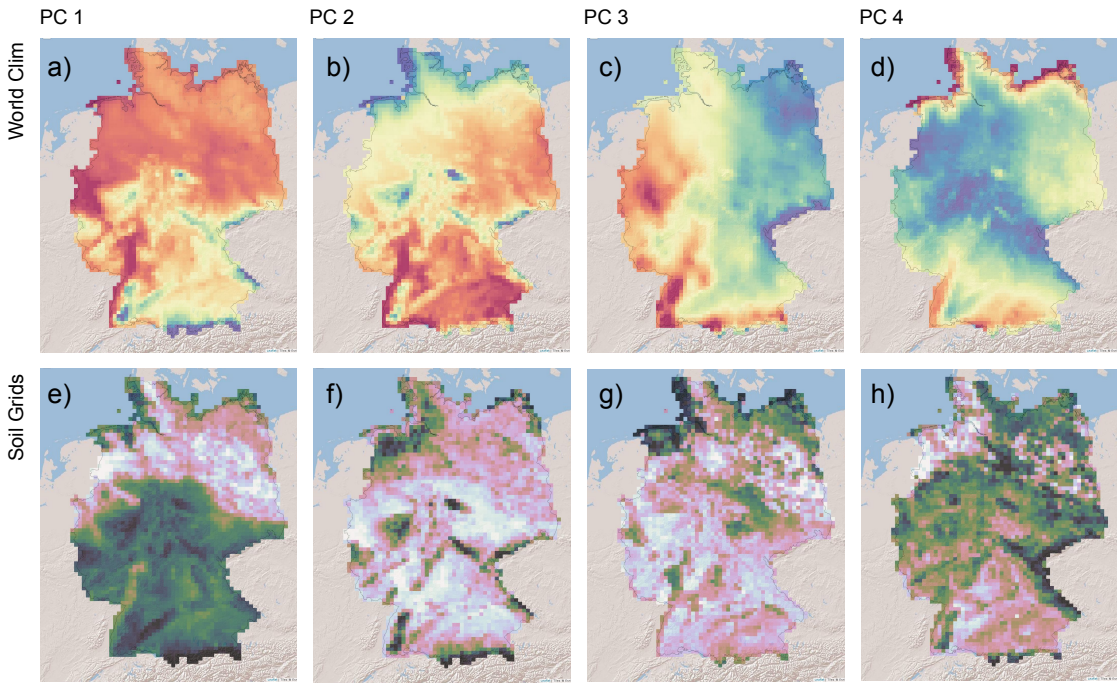

Figure S8: Leading principal components used for prediction in geographical space.

fraction per se (SNDPPT).

- $\text{PC}_{\text{Soil}}^{(2)}$ = "Organic carbon axis". On this axis the highest contribution comes from Soil Organic Carbon Content (ORCDRC) as well as Bulk Density (BLDFIE),

- $\text{PC}_{\text{Soil}}^{(3)}$ = "pH axis". Here, we clearly have all kind of variables related to soil pH (e.g. PHIKCL).

- $\text{PC}_{\text{Soil}}^{(4)}$ = This is yet a second water related axis, which is however more related to soil clay content.

1.2 Parameter selection for dimensionality reduction

Nonlinear dimensionality reduction via Isomap (Tenenbaum et al., 2000) requires a distance matrix. In the case of binary data we can use the Jaccard distances that then form the basis for computing a k-NN graph. Dijsktra-shortest path algorithm is then used to find the geodesic distances along this graph. Note for R-users that we presume there is a bug in the vegan-package and one should use alternatives e.g. a graph computing package to compute correct geodesic distances. All code to reproduce the paper is provided in the accompanying zip file. In fig. S9 we show the explained variances for Isomap based on varying the k-NN parameter.

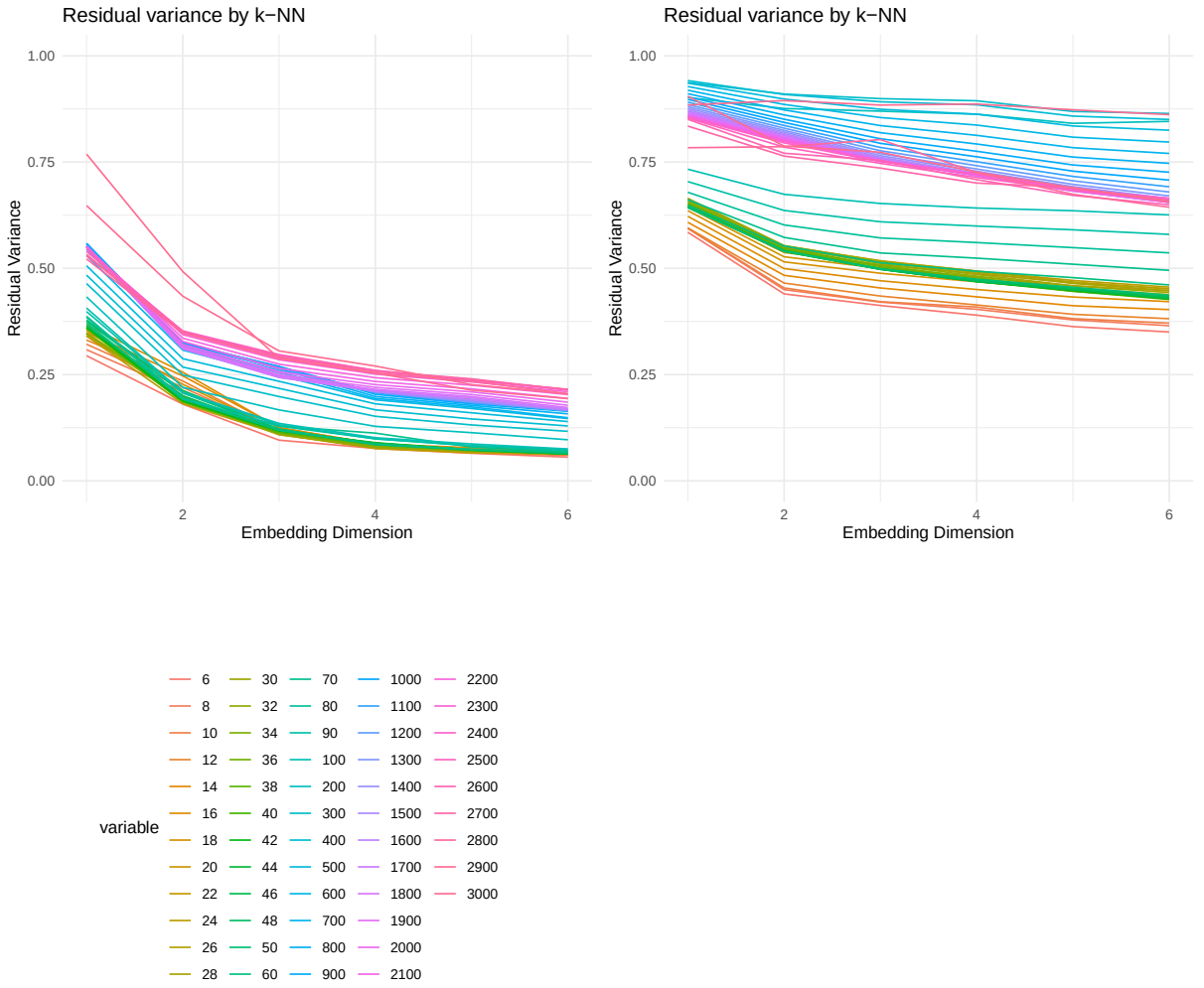

Figure S9: Residual variances per embedding dimension. The dimensionality reduction was performed via Isomap for varying k-values starting from $k = 6$ until $k = 30$, with increasing step length. Left: results for Florkart; right: results for Flora Incognita. For Florkart the compression converges for four dimensions if k is sufficiently small; the exact choice of k is not that relevant and any k <50 is acceptable, for Flora Incognita, there is a stability between of $k < 40$ after which the compression decreases. Smallest k-values achieve more compression though, i.e. in the higher dimensions.

1.3 Canonical correlation analysis, CCorA

The output of the Isomap analysis of Florkart and Flora Incognita are Isomap dimensions. Joint patterns can be identified via CCorA, i.e. we find linear combinations of all Isomap dimensions of Florkart that maximize the correlation to the Isomap dimensions estimated from Flora Incognita.

From fig. S9 we know that small k-values favour data compression. However, we later want to analyse the joint patterns. In Fig. S10 we performed a CCorA among all combinations to embeddings (varying k-value) and report the sum of the canonical correlations for the leading four canonical variates (which we tested to be significant in most combinations). The aim is to find the smallest k-NN value that has a maximal value in this figure. This value is clearly $k = 16$. All subsequent analysis are based on this embedding setting for both data sets. All subsequent dimensionality reduction analyses will be based on an Isomap embedding considering $k = 16$.

In fig. S11 we show the weights of the respective Isomap ($k = 16$) dimensions on the canonical variates.

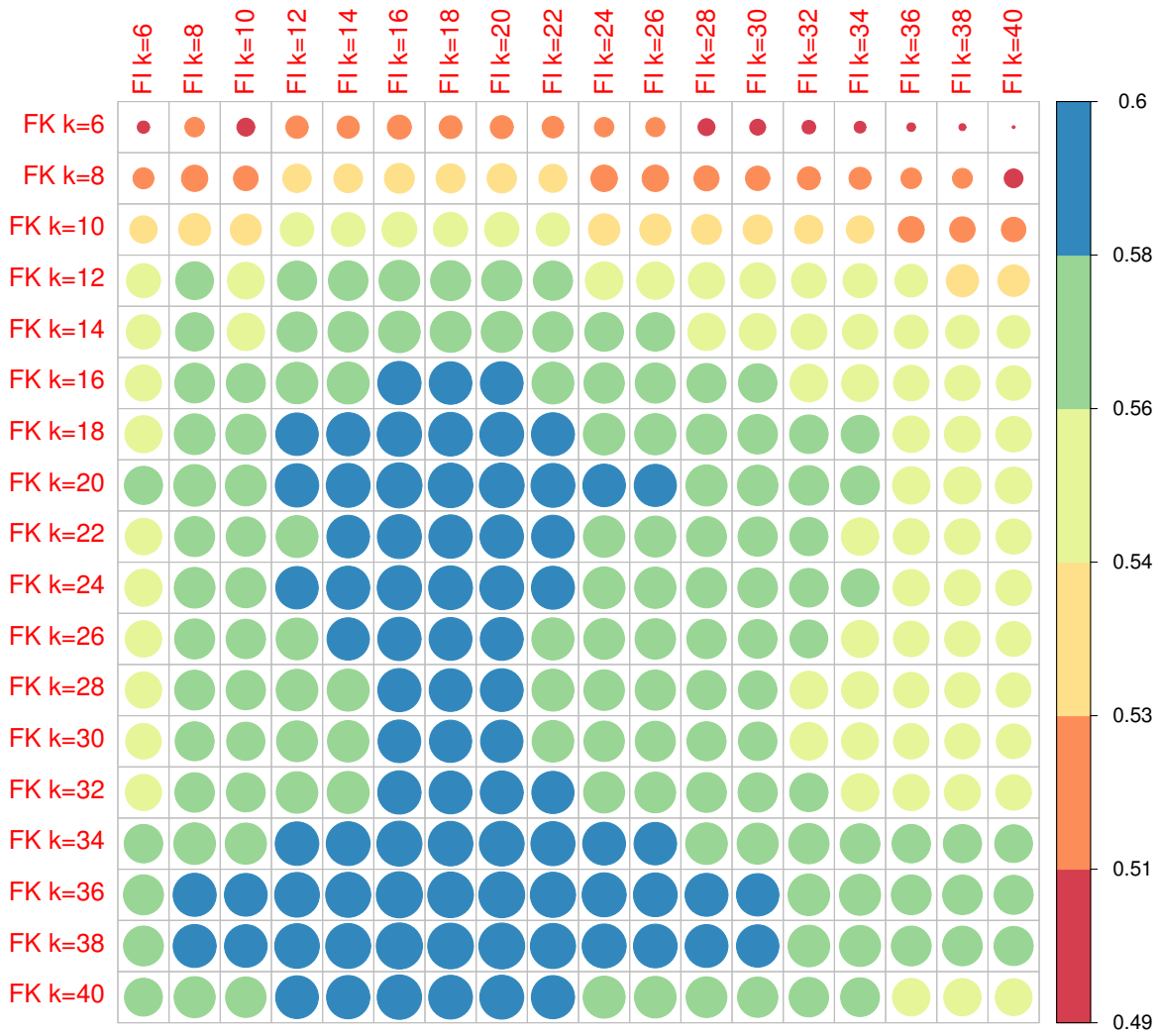

Figure S10: Sum of the canonical correlations for the leading four canonical variates for varying k-NN values in the embedding of Florkart (FK) and Flora Incognita (FI). Circle sizes scale with colorbar.

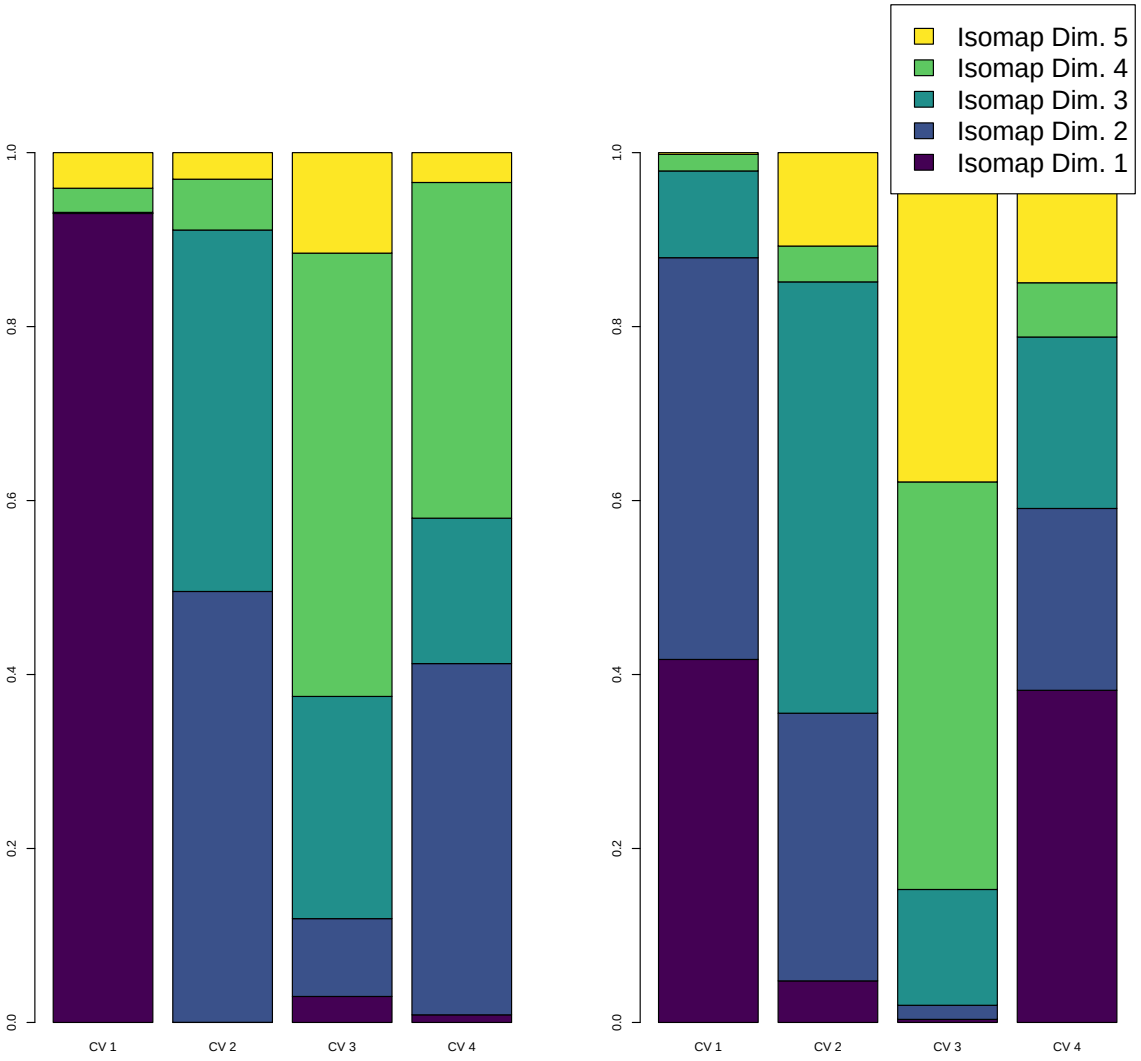

Figure S11: Contributions of the different Isomap components to the canonical variates. Left: Florkart; right: Flora Incognita. In case of Florkart, the first canonical variate is exclusively composed from the first Isomap dimension, while in the case of Flora Incognita patterns of the first two dimension are combined to almost equal fractions. I). Circle sizes scale with colorbar.

1.4 Prediction results of the spatial patterns

We used the principal components as described above as potential predictors in random forests variable selection, considering spatial cross validation. For defining spatial folds for the cross validation we clustered the leading Isomap dimensions combined in an asymmetric similarity matrix with geographical proximity using affinity propagation (described in detail in the main text Frey and Dueck, 2007). The spatial folds are shown in fig. S11. The results of the variable selection and importance are summarized in fig. S12 (as Fig. 4 i-l but here seperated by canonical variates).

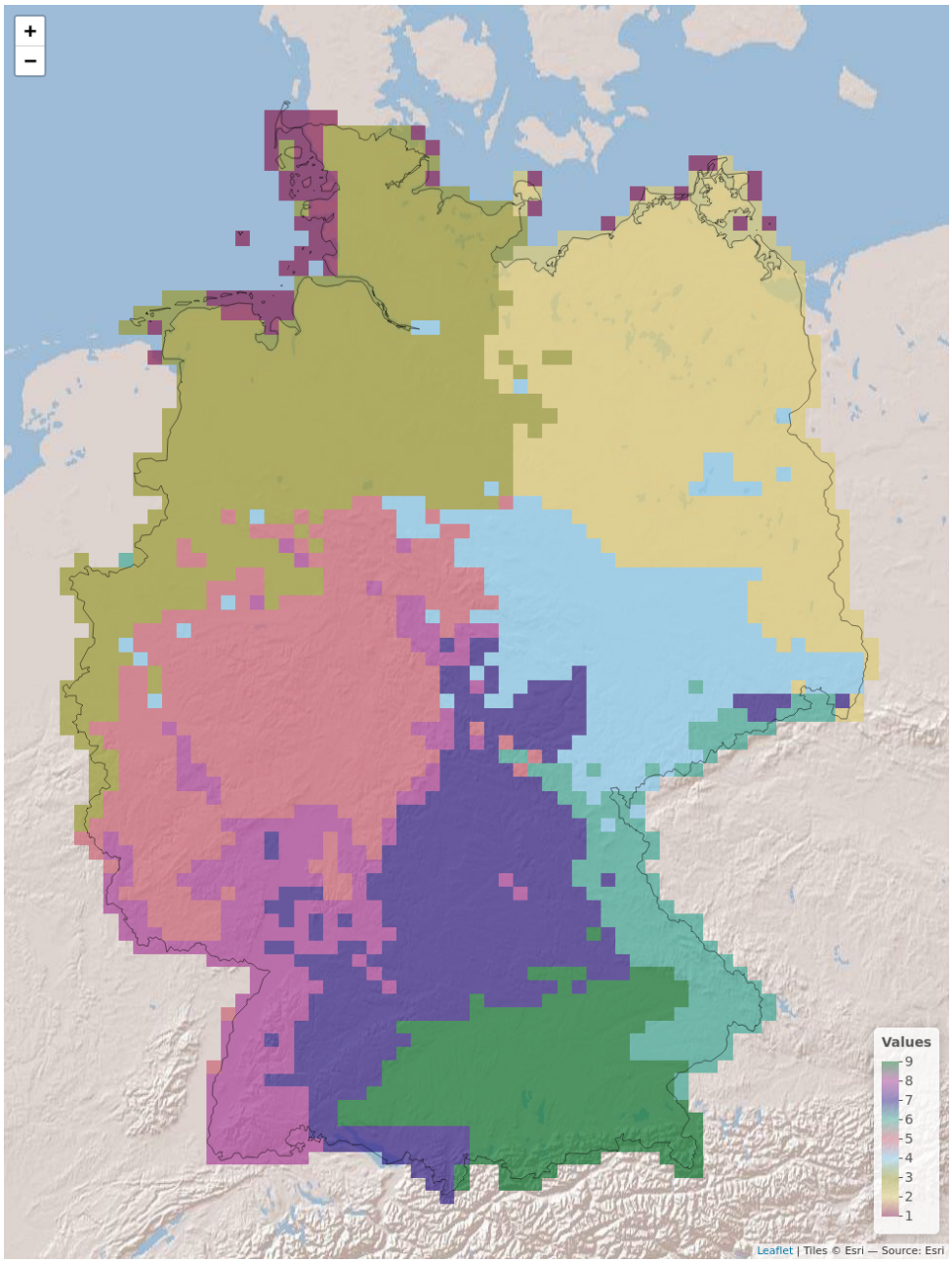

Figure S12: Results of the clustering of the leading Isomap coordinates (of Florkart) using the "Affinity Propagation" based on an asymmetric similarity matrix. While the one triangle contained inverse Euclidean distances, the other was based on geographical vicinity. Hence, the similarity of two grid cells in one direction was estimated using their floristic similarity, while the inverse message estimated geographical distance. This approach guarantees almost connected spatial clusters. The emerging cluster number was 9.

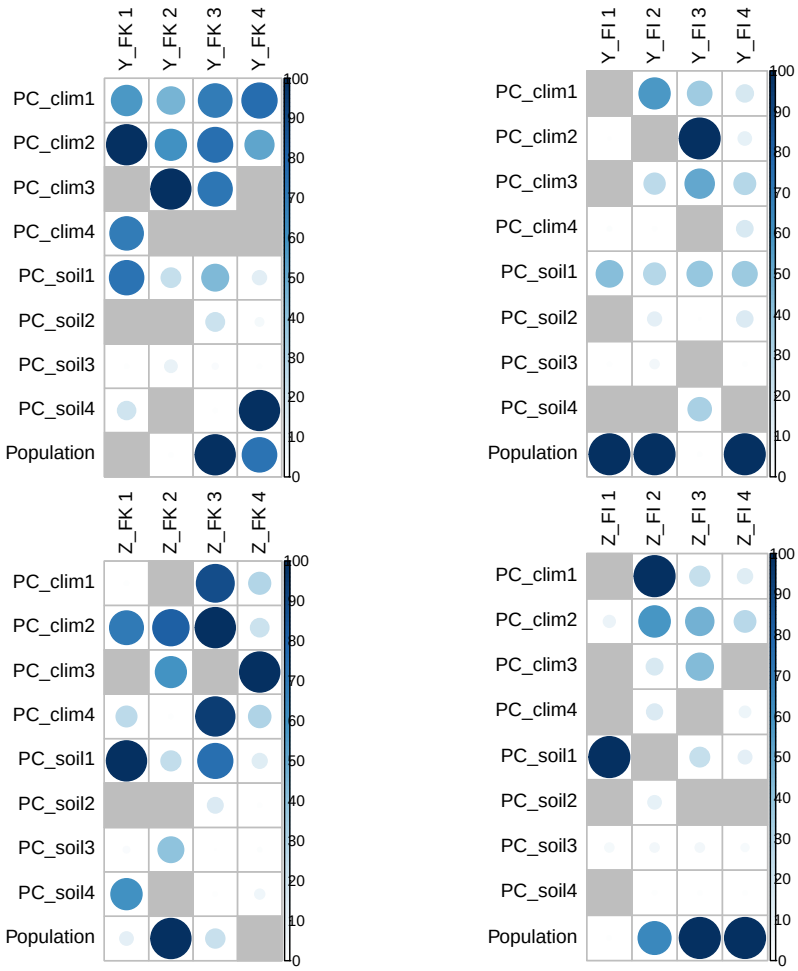

Figure S13: Relative importances of variables for predicting the different potential targets after cross validation considering spatial folds. Upper left: Selected variables for the Isomap dimensions of Forkart $\mathbf{y}_{\text{FK}}^{(i)}$, $i \in \{1, \ldots, 4\}$; upper right: analogous for the Flora Incognita dimensions $\mathbf{y}_{\text{FI}}^{(i)}$. Lower left: Canonical variates for Florkart $\mathbf{z}_{\text{FK}}^{(i)}$, lower right: Canonical variates for Flora Incognita $\mathbf{z}_{\text{FI}}^{(i)}$. Gray values show cases where a predictor was not selected for the given target variable.

References

Fick, S. E., and R. J. Hijmans. 2017. Worldclim 2: new 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology* **37**, 4302–4315.

Frey, B. J., and D. Dueck. 2007. Clustering by passing messages between data points. *Science* **315**, 972–976.

Hengl, T., J. M. de Jesus, R. A. MacMillan, N. H. Batjes, G. B. Heuvelink, E. Ribeiro, A. Samuel-Rosa, B. Kempen, J. G. Leenaars, M. G. Walsh *et al.* 2014. Soilgrids1km – global soil information based on automated mapping. *PloS one* **9**, e105992.

Tenenbaum, J. B., V. De Silva, and J. C. Langford. 2000. A global geometric framework for nonlinear dimensionality reduction. *Science* **290**, 2319–2323.