**1. Project**

**Title:** The SPECIAL Modern Pollen Data Set for Climate Reconstructions, version 2 (SMPDSv2)

**Dates:** 2022

**Funding organization**: Funding for the development of the SPECIAL Modern Pollen Data Set for Climate Reconstructions, version 2 (SMPDSv2) was sourced from an ERC-funded project GC2.0 (Global Change 2.0: Unlocking the past for a clearer future, grant: ERC 694481)

**2. Dataset**

**Title:** The SPECIAL Modern Pollen Data Set for Climate Reconstructions, version 2 (SMPDSv2)

**Summary description:** The dataset contains counts for the [unique\_pollen\_taxa] pollen taxa from individual modern samples from around the world and supporting metadata about each sample. The dataset has been specifically designed for use to make quantitative reconstructions of climate and vegetation parameters.

**Publication year:** 2022

**Creators:** Roberto Villegas-Diaz and Sandy P. Harrison

**Organisation:** Geography and Environmental Science, University of Reading, UK

**Rights Holder:** University of Reading

**3. Terms of use**

This dataset is licensed by the rights-holder(s) under a Creative Commons Attribution 4.0 International Licence: <https://creativecommons.org/licenses/by/4.0/>.

In order to assure traceability, any presentation, report, or publication that uses the SMPDSv2 should cite the dataset (https://doi.org/10.17864/XYZ)

**4. Contents**

Pollen data are widely used to reconstruct past climate changes, using relationships between modern pollen abundance in surface samples and climate at the surface sample sites. These records can also be used in a similar way to reconstruct past vegetation. The quality of the climate and vegetation reconstructions is strongly influenced by whether the training data set provides an adequate sampling of the climate space. We have assembled modern pollen records from [unique\_number\_of\_sites] sites worldwide, where modern is defined as post 50 cal. yr BP. For pollen trap and surface samples, the age was taken as the year of collection unless otherwise specified by the author. The ages of samples from cores or sections are based on the constructed (original) age model; multiple samples may be included from high-resolution records if they post-date 50 cal. yr BP.

The pollen data were compiled from multiple different published regional datasets, from data repositories (Neotoma, PANGAEA) or directly from data collectors/authors. Several of the regional data sets cover overlapping geographical areas and include some of the same sites; some of these data sets were created by combining smaller data sets put together by different people and themselves contain duplicated sites with slightly different metadata. We have privileged original records from repositories over data compilations. We then screened each of the data sets to remove any sites duplicated within the dataset or in other data sets, on the basis of comparison of the metadata and the pollen counts. In cases where there was still ambiguity, we contacted the original data providers for clarification. Site type, basin size and entity type provide important information for selecting samples for quantitative reconstructions. Although some of the datasets provided this information, the descriptions were freeform and not standardised; other source did not provide these metadata. We have standardised the descriptions of site and entity type, and where this information was lacking we consulted the original publications for each site.

Non-pollen palynomorphs (e.g. charcoal, spores, algae) were removed from the data set. Obligate aquatics, insectivorous plants, and cultivated plants were also removed from the data set, since their distribution is not primarily controlled by climate. Most of the records are raw counts, but some records were provided as percentages. No attempt was made to reconstruct the counts from these percentages because the basis of the pollen sum was not specified in every case. Given the mix of counts and percentages, the data from individual records need to be normalised before use.

The original taxonomy was standardised using Plants of the World Online ([www.plantsoftheworldonline.org/](http://www.plantsoftheworldonline.org/)) and the Integrated Taxonomic Information System (<https://www.itis.gov/>). This "CLEAN" dataset consists of [unique\_pollen\_taxa] taxa (Table 1). The level of taxonomic identification varies across the data sets, between regions and between individual palynologists. The reliability of climate reconstructions can be compromised if pollen taxa are not routinely identified across all the sites in a training data set because of uneven sampling of the climate space. Furthermore, some of the clean pollen types are rare and were recorded at too few sites to allow the construction of robust relationships with climate. We therefore provide two date sets with the pollen combined to a higher taxonomic level. In the first ("INTERMEDIATE"), we preserved species-level identifications for trees, small trees and widespread erect shrubs but combined other taxa to genus level (Table 1). In the second ("AMALGAMATED"), woody species are generally combined at genus level and herbaceous species at sub-family or family level unless they are palaynologically distinctive and geographically widespread. The "AMALGAMATED" dataset consists of [unique\_pollen\_taxa] taxa (Table 1).

To facilitate the use of these data for climate and vegetation reconstructions, we provide estimates of the climate and the potential natural vegetation at each terrestrial site. We do not estimate these for marine records or from very large lakes (e.g. Caspian Sea) because the source area for such records is extremely large. Modern climate data were derived from the Climate Research Unit CRU TS4.04 data set, which provides monthly mean precipitation, temperature and fractional sunshine hours for the period 1961 to 1990 at 0.5 degrees spatial resolution (Harris et al., 2020). We used geographically weighted regression using latitude, longitude and elevation as predictors to obtain the climate at the location and elevation of each pollen site and calculated mean annual precipitation (MAP) and mean annual temperature (MAT) at each site. We then calculated bioclimatic variables using these data, specifically mean temperature of the coldest month (MTCO), mean temperature of the warmest month (MTWA) growing degree says above a baseline of 0°C (gdd0), and a moisture index (MI) calculated as the ratio of annual precipitation to the estimated annual equilibrium evapotranspiration, using the Simple Process-Led Algorithms for Simulating Habitats (SPLASH) model (Davis et al., 2017). The climate variables were chosen because they reflect distinctive physiological controls on plant growth. The modern potential natural vegetation (PNV) at each pollen site was extracted from an updated version of the Global PNV map produced by Hengl et al. (2018). The original version of this map had a resolution of 1 km; the updated version has a resolution of 250m. See Appendix 1 for a step by step of the derivation of the climate variables.

**File structure:** The data are stored in a relational database (MySQL), which consists of 4 linked tables, specifically: "entity", "climate", "taxon\_name" and "pollen\_count" (Figure 1). The database is also stored as 4 flat CSV files:

* "smpdsv2\_metadata.csv": combines the "entity" and "climate" tables
* "smpdsv2\_pollen\_counts\_clean.csv": contains the clean version of the pollen counts, where each row represents the samples and the columns are: ID\_SAMPLE, taxon1, taxon2, …, taxonk
* "smpdsv2\_pollen\_counts\_intermediate.csv": contains the intermediate version of the pollen counts, where each row represents the samples and the columns are: ID\_SAMPLE, taxon1, taxon2, …, taxonm
* "smpdsv2\_pollen\_counts\_amalgamated.csv": contains the amalgamated version of the pollen counts, where each row represents the samples and the columns are: ID\_SAMPLE, taxon1, taxon2, …, taxonn

As these are flat CSV files, no relationships are defined here but the tables can be joined using different programming languages (R, Python, etc.) based on the foreign keys (shared column names between tables such as ID\_ SAMPLE in the metadata and files with pollen counts). The different fields included in the database are summarised in Table 2.

Additionally, a companion R data package has been created and can be found at <https://github.com/special-uor/smpds>. This package contains the 4 tables in the database, as well as a set of utilitarian functions to generate plots and create climate reconstructions.

**Access to the SMPDSv2:** The SMPDSv2 is stored as a MySQL database file ("smpdsv2.sql"). Please check https://dev.mysql.com/downloads/ to download and install MySQL. Once MySQL Community Server and MySQL Workbench (or any other database client of your preference) are installed, the database can be imported and visualised. A schema must be created upon import. To import the SQL file, you follow:

1. Open MySQL Workbench

2. Connect to the connection you would like to store your database in. A connection is usually created during the installation process (usually root@localhost with the password defined during the installation process)

3. Server > Data Import > Import from Self-contained file

4. Browse to the SQL file you have downloaded

5. Press New, next to the default target schema to create a new schema (name this as appropriate: e.g. SMPDSv2)

6. Press Import

**5. References**

Davis, B.A., Chevalier, M., Sommer, P., Carter, V.A., Finsinger, W., Mauri, A., Phelps, L.N., Zanon, M., Abegglen, R., Åkesson, C.M. and Alba-Sánchez, F., 2020. The Eurasian Modern Pollen Database (EMPD), version 2. *Earth System Science Data*, 12(4), pp.2423-2445, <https://doi.org/10.5194/essd-12-2423-2020>

Gaillard, M.J., Birks, H.J.B., Emanuelsson, U. and Berglund, B.E., 1992. Modern pollen/land-use relationships as an aid in the reconstruction of past land-uses and cultural landscapes: an example from south Sweden. *Vegetation History and Archaeobotany*, *1*(1), pp.3-17, <https://doi.org/10.1007/BF00190697>

Hai-Yan, C., De-Yu, X.U., Meng-Na, L., Kai, L.I., Jian, N.I., Xian-Yong, C., Bo, C., Xiu-Dong, H., Zhao-Chen, K., Sheng-Feng, L., Xiao-Qiang, L., Guang-Xiu, L., Ping-Mei, L., Xing-Qi, L., Xiang-Jun, S., Ling-Yu, T., Hai-Cheng, W., Qing-Hai, X., Shun, Y., Xiang-Dong, Y., Zhen-Jing, Y., Ge, Y.U., Yun, Z., Zhi-Yong, Z., Ke-Liang, Z., Zhuo, Z., Herzschuh, U., 2021. A modern pollen dataset of China. Chinese Journal of Plant Ecology 45, 799. <https://doi.org/10.17521/cjpe.2021.0024>

Lézine, A.-M., 1988. Les variations de la couverture forestière mésophile d’Afrique occidentale au cours de l’Holocène. C. r. Acad. sci., Sér. 2, Méc. phys. chim. sci. univers sci. terre 307, 439–445.

Harris, I., Osborn, T.J., Jones, P. and Lister, D., 2020. Version 4 of the CRU TS monthly high-resolution gridded multivariate climate dataset. *Scientific Data*, *7*(1), pp.1-18, <https://doi.org/10.6084/m9.figshare.11980500>

Harrison, S.P., 2019. Modern pollen data for climate reconstructions, version 1 (SMPDS). Dataset. <https://doi.org/10.17864/1947.194>

Harrison, S.P., Marinova, E. and Cruz-Silva, E., 2021. EMBSeCBIO pollen database. Dataset. <https://doi.org/10.17864/1947.309>

Harrison, S. P., Shen, Y. and Sweeney, L., 2022. Pollen data and charcoal data of the Iberian Peninsula (version 3). Dataset. <https://doi.org/10.17864/1947.000369>

Hengl, T., Walsh, M.G., Sanderman, J., Wheeler, I., Harrison, S.P. and Prentice, I.C., 2018. Global mapping of potential natural vegetation: an assessment of machine learning algorithms for estimating land potential. *PeerJ*, *6*, p.e5457, <https://doi.org/10.7717/peerj.5457>

Herzschuh, U., Cao, X., Laepple, T., Dallmeyer, A., Telford, R.J., Ni, J., Chen, F., Kong, Z., Liu, G., Liu, K.B. and Liu, X., 2019. Position and orientation of the westerly jet determined Holocene rainfall patterns in China. *Nature Communications*, 10(1), pp.1-8, <https://doi.org/10.1038/s41467-019-09866-8>

Phelps, L.N., Chevalier, M., Shanahan, T.M., Aleman, J.C., Courtney‐Mustaphi, C., Kiahtipes, C.A., Broennimann, O., Marchant, R., Shekeine, J., Quick, L.J. and Davis, B.A., 2020. Asymmetric response of forest and grassy biomes to climate variability across the African Humid Period: influenced by anthropogenic disturbance?. *Ecography*, 43(8), pp.1118-1142, <https://doi.org/10.1111/ecog.04990>

Vincens, A., Lézine, A.M., Buchet, G., Lewden, D. and Le Thomas, A., 2007. African pollen database inventory of tree and shrub pollen types. *Review of Palaeobotany and Palynology*, 145(1-2), pp.135-141, <https://doi.org/10.1016/j.revpalbo.2006.09.004>

Williams, J.W., Grimm, E.C., Blois, J.L., Charles, D.F., Davis, E.B., Goring, S.J., Graham, R.W., Smith, A.J., Anderson, M., Arroyo-Cabrales, J. and Ashworth, A.C., 2018. The Neotoma Paleoecology Database, a multiproxy, international, community-curated data resource. *Quaternary Research*, 89(1), pp.156-177, <https://doi.org/10.1017/qua.2017.105>

A picture containing text

Description automatically generated

*Figure 1. Entity-relation diagram showing the structure of the database, each individual table, their attributes, and relationships are shown. One-to-many relations indicate that is possible for several entities in one table to be linked to a single record in another table (e.g. the same taxon\_name can be represented in many pollen\_count records, for different samples). The database uses foreign keys (FK) to allow for such linkages.*

*Table 2. Summary of the different fields in the SMPDSv2 and the tables in which they are found.*

|  |  |  |  |
| --- | --- | --- | --- |
| **Table** | **Field name** | **Definition** | **Format** |
| entity | ID\_SITE | Unique identifier for each site. A site can have multiple entities. | Unsigned integer |
| entity | ID\_ENTITY | Unique identifier for each entity. An entity can have multiple samples at different depths. | Unsigned integer |
| entity, climate, pollen\_count | ID\_SAMPLE | Unique identifier for each sample | Unsigned integer |
| entity | source | Source from which the data was repatriated (e.g. Neotoma, EMPDv2) | Text |
| entity | site\_name | Site name as given by original authors or as defined by us where there was no unique name given to the site | Text |
| entity | entity\_name | Name of entity, where an entity may be a separate core from the same site or an alternative measurement type of the same core. | Text |
| entity | latitude | Latitude of the sampling site, given in decimal degrees, where N is positive and S is negative | Double precision number |
| entity | longitude | Longitude of the sampling site in decimal degrees, where E is positive and W is negative | Double precision number |
| entity | elevation | Elevation of the sampling site in metres above (+) or below (-) sea level | Double precision number |
| entity | basin\_size | Size of sampled site (e.g. lake, bog, etc.) in km2 or given as a categorical estimated when precise information was not recorded or was not available | Text |
| entity | site\_type | Information about type of site (e.g. cave, lake, glacial, terrestrial, etc.) | Text |
| entity | entity\_type | Information about the type of entity (e.g. composite, core top, surface sample, etc.) | Text |
| entity | age\_BP | Sample age in years before present (BP) or categorical estimate where not numeric values were available | Text |
| entity | publication | Citation for the publication where the data was originally published | Text |
| entity | doi | Digital Object Identifier (DOI) for the publication | Text |
| climate | ID\_BIOME | Unique identifier for each potential natural vegetation (PNV) | Unsigned integer |
| climate | PNV | Potential Natural Vegetation based on the work by Hengl et al., 2018 | Text |
| climate | mi | Reconstructed Moisture Index [unitless] using the CRU TS 4.04 (Harris et al., 2020) data set | Double precision number |
| climate | gdd0 | Reconstructed Growing Degree Days above 0 ºC [ºC days] using the CRU TS 4.04 (Harris et al., 2020) data set | Double precision number |
| climate | mat | Reconstructed Mean Annual Temperature [ºC] using the CRU TS 4.04 (Harris et al., 2020) data set | Double precision number |
| climate | mtco | Reconstructed Mean Temperature of the Coldest Month [ºC] using the CRU TS 4.04 (Harris et al., 2020) data set | Double precision number |
| climate | mtwa | Reconstructed Mean Temperature of the Warmest Month [ºC] using the CRU TS 4.04 (Harris et al., 2020) data set | Double precision number |
| climate | map | Reconstructed Mean Annual Precipitation [mm/year] using the CRU TS 4.04 (Harris et al., 2020) data set | Double precision number |
| taxon\_name,  pollen\_count | ID\_TAXON | Unique identifier for each taxon | Unsigned integer |
| taxon\_name | taxon\_name | Standardised taxon name | Text |
| pollen\_count | amalgamation\_level | Level of amalgamation for the pollen counts:   * amalgamation\_level = 0:   clean pollen counts   * amalgamation\_level = 1:   intermediate pollen counts   * amalgamation\_level = 2:   amalgamated pollen counts | Unsigned integer between 0 and 2 |
| pollen\_count | count | Standardised pollen count | Double precision number |

Table 1: List of pollen taxa.

|  |
| --- |
| Abies |

[to be included after SPH’s inspections are implemented] This will need to include the clean, intermediate and amalganated names

Table Z: List of data sources and references

|  |  |  |
| --- | --- | --- |
| source (metadata table) | Number of entities | Publications |
| AMSS | 38 | (Jolly et al., 1996; Julier et al., 2019, 2018; Lebamba et al., 2009) |
| APD | 90 | (Vincens et al., 2007) |
| Australian pollen | 1540 | (Adeleye et al., 2021b, 2021a; Beck et al., 2017; Field et al., 2018, 2018; Fletcher et al., 2014; Herbert and Harrison, 2016; Luly, 1993; Luly et al., 1986; Mariani et al., 2017; McWethy et al., 2014, 2010; Pickett et al., 2004; Prebble et al., 2019) |
| BIOME6000 Japan | 94 | (Takahara et al., 2000) |
| Blyakharchuk | 144 | Author: Tatiana A. Blyakharchuk |
| Bush et al., 2021 | 636 | (Bush et al., 2021) |
| CMPD | 4208 | (Hai-Yan et al., 2021) |
| Dugerdil et al., 2021 | 48 | (Dugerdil et al., 2021) |
| EMBSeCBIO | 149 | (Harrison et al., 2021) |
| EMPDv2 | 3508 | (Davis et al., 2020) |
| Gaillard et al., 1992 | 124 | (Gaillard et al., 1992) |
| Harrison et al., 2021 | 3 |  |
| Herzschuh et al., 2019 | 595 | (Herzschuh et al., 2019) |
| IBERIA | 243 | (Harrison et al., 2022) |
| Neotoma | 6702 | (Williams et al., 2018) |
| Phelps et al., 2020 | 106 | (Phelps et al., 2020) |
| SMPDSv1 | 6345 | (Harrison, 2019) |
| Southern Hemisphere pollen | 76 | (Black, 2006; Dodson, 1978; Dodson and Intoh, 1999; Haberle, 1996, 1993; Hope, 2009; Hope et al., 1999, 1988; Macphail, 1980, 1979, 1975; Macphail and McQueen, 1983; Macphail and Mildenhall, 1980; Norton et al., 1986; Prebble et al., 2010; Shulmeister et al., 2003) |