1. Project

Title: Eastern Mediterranean-Black Sea-Caspian-Corridor Biomes (EMBSeCBIO) project

Dates: September 2007 – ongoing

Funding organisations: European Research Council (ERC)

Grant number: ERC 694481_GC2.0

2. Dataset

Title: EMBSeCBIO pollen database

Summary information: The EMBSeCBIO pollen database is a compilation of pollen counts or pollen percentages for 1132 modern entities (modern is defined as younger than 150 cal. years) and 187 fossil entities, in the Mediterranean-Black Sea Caspian-Corridor, located between 28°-49°N and 20°-62°E. The database includes tables describing the characteristics of the sites from which the records were obtained. Information on dating and the original age-depth models for the fossil records are included. New age-depth models have been created using the IntCal20 calibration curve for 148 records.

Publication year: 2021

Creators: Sandy P. Harrison, Elena Marinova and Esmeralda Cruz-Silva

Organisations: University of Reading and University of Leuven

3. Terms of use

4. Contents

Pollen records provide information to reconstruct past changes in vegetation and climate. The Mediterranean-Black Sea Caspian-Corridor, located between 28°-49°N and 20°-62°E, provides an ideal natural laboratory to examine millennial vegetation dynamics and their potential connexions with climate events at sub-continental scale. This region is characterized by strong temperature and precipitation gradients, topographic heterogeneity, and more than 30 000 years of human occupation (Cordova et al., 2009).

The compilation of the pollen records of the Mediterranean-Black Sea Caspian Corridor in a single database (the EMBSeCBIO database) was started in 2009 as a project under the auspices of the Palynology Working Group (WG-2) UNESCO-sponsored International Geoscience Programme IGCP-521 (Cordova et al., 2009), and it has been continued over time (Marinova et al., 2018). The current release (EMBSeCBIO pollen database) is the compilation of the modern pollen data and the fossil pollen records for the region. The EMBSeCBIO pollen database contains pollen data for individual records (entities) grouped by sites. Additional tables provide information on dating, including information on the dates used to construct the age models. The metadata tables provide information about the characteristics of the sites from which the records were obtained. Missing information, including dating information, has been added to the database and some records have been amended where mistakes were found. New age-depth models have been created using the IntCal20 calibration curve (Reimer et al., 2020) and the rbacon R package (Blaauw et al., 2021) in the framework of the AgeR R package (Villegas-Diaz et al., 2021) for 148 entities.

4.1 Description of files

EMBSeCBIO_pollen_DB.sql

There is a single MySQL database file (EMBSEcBIO_pollen_DB.sql). Please check https://dev.mysql.com/downloads/ to download and install MySQL. Once MySQL Community Server and MySQL Workbench 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 option button, next to the Default Target Schema, to create a new schema (name this as appropriate, such as EMBSeCBIO)
6) Press Import

Please note that once the database is imported, there are packages and modules in several programming languages which will allow you to connect to the database such as RMySQL in R, and MySQLdb in python.

EMBSeCBIO_pollen_DB.zip

There is a single compressed archive file (EMBSeCBIO_pollen_DB.zip) comprising 15 CSV files corresponding to the 15 individual tables in the MySQL database. The CSV file names correspond to the table names. As these are flat CSV files, no relationships are defined here but the tables can be joined in different programming languages (R, Python, etc.) based on the foreign keys (shared column names between tables such as ID_SITE in the site and entity tables). The relationships are described in figure 1 and the characteristics of each table are described in tables 1 to 15. Please note that CSV files are in UTF-8 characters, and special characters (such as Greek characters, and letters with accents which may appear in site names and in citations) may not be reproduced correctly when open as default in Excel.

Therefore, due to the multilingual nature of the site/entity names, you will need to follow these steps to open the csv data files with Excel in Windows computers (otherwise the UTF-8 encoding is not recognised):

1) Open Excel
2) Import the data using Data -> Import External Data--> Import Data
3) Select the file type of "csv" and browse to your file
4) In the import wizard change the File_Origin to "65001 UTF-8"
5) Change the Delimiter to comma
6) Select where to import to and Finish

EMBSeCBIO_pollen_DB_codes.zip

There is a single compressed archive file (EMBSeCBIO_pollen_DB_codes.zip) comprising examples of codes and queries that can be used with the MySQL database, but also with the CSV file. Within this compressed file there is:

- An html file (EMBSeCBIO_DB_query_example.html) which show examples of SQL queries on the database
- An R file (EMBSeCBIO_connectDB.R) demonstrating how to connect R to the database once the database has been uploaded into MySQL.

Please note that there may be some authentication issues when using MySQL 8.0, especially when trying to connect from R/Python. This could be due to the change in the default authentication plugin from mysql_native_password to caching_sha2_password. One way around this is to run the following MySQL query in MySQL Workbench:

ALTER USER 'username’@‘host’ IDENTIFIED WITH mysql_native_password BY ‘password’;

where ‘username’ refers to the user’s username (‘root’ if MySQL is run locally), ‘host’ refers to the host name (‘localhost’ if MySQL is run locally) and ‘password’ refers to the password (if MySQL is run locally, this is usually the password set up when installing MySQL).

5. References

Blaauw, M., Christen, J. A., Lopez, M. A. A., Vazquez, J. E., V, O. M. G., Belding, T., Theiler, J., Gough, B., & Karney, C. (2021). *rbacon: Age-Depth Modelling using Bayesian Statistics* (2.5.6) [Computer software]. https://CRAN.R-project.org/package=rbacon

Blaauw, M., Christen, J. A., Lopez, M. A. A., Vazquez, J. E., V, O. M. G., Belding, T., Theiler, J., Gough, B., & Karney, C. (2021). *rbacon: Age-Depth Modelling using Bayesian Statistics* (2.5.6) [Computer software]. https://CRAN.R-project.org/package=rbacon

Cordova, C. E., Harrison, S. P., Mudie, P. J., Riehl, S., Leroy, S. A. G., & Ortiz, N. (2009). Pollen, plant macrofossil and charcoal records for palaeovegetation reconstruction in the Mediterranean-Black Sea Corridor since the Last Glacial Maximum. *Quaternary International*, *197*(1–2), 12–26. https://doi.org/10.1016/j.quaint.2007.06.015

Harrison, S. P., & Marinova, E. (2017). *EMBSeCBIO modern pollen biomisation* [Data set]. University of Reading. https://doi.org/10.17864/1947.109

Reimer, P. J., Austin, W. E. N., Bard, E., Bayliss, A., Blackwell, P. G., Ramsey, C. B., Butzin, M., Cheng, H., Edwards, R. L., Friedrich, M., Grootes, P. M., Guilderson, T. P., Hajdas, I., Heaton, T. J., Hogg, A. G., Hughen, K. A., Kromer, B., Manning, S. W., Muscheler, R., … Talamo, S. (2020). The IntCal20 Northern Hemisphere Radiocarbon Age Calibration Curve (0–55 cal kBP). *Radiocarbon*, *62*(4), 725–757. https://doi.org/10.1017/RDC.2020.41

Villegas-Diaz, R., Cruz-Silva, E., & Harrison, S. P. (2021). *ageR: Supervised Age Models*. Zenodo. https://doi.org/10.5281/zenodo.4636716

Reimer, P. J., Austin, W. E. N., Bard, E., Bayliss, A., Blackwell, P. G., Ramsey, C. B., Butzin, M., Cheng, H., Edwards, R. L., Friedrich, M., Grootes, P. M., Guilderson, T. P., Hajdas, I., Heaton, T. J., Hogg, A. G., Hughen, K. A., Kromer, B., Manning, S. W., Muscheler, R., … Talamo, S. (2020). The IntCal20

Northern Hemisphere Radiocarbon Age Calibration Curve (0–55 cal kBP). *Radiocarbon*, *62*(4), 725–757. https://doi.org/10.1017/RDC.2020.41

Villegas-Diaz, R., Cruz-Silva, E., & Harrison, S. P. (2021). *ageR: Supervised Age Models*. Zenodo. https://doi.org/10.5281/zenodo.4636716

6. Figures and Tables

Figure 1. The structure of the EMBSeCBIO_pollen_DB database

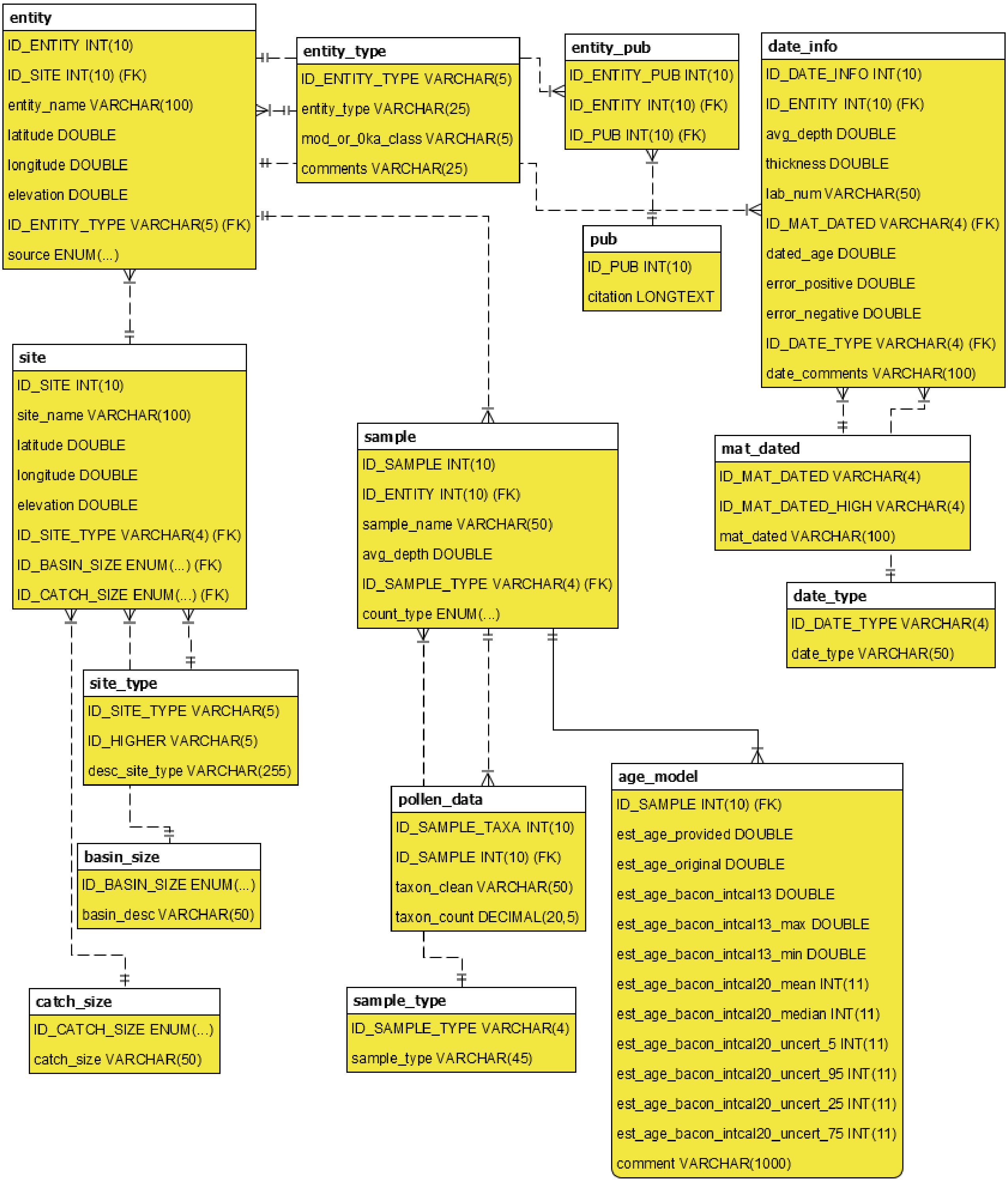

Table 1. Characteristics of the **site** table

Field label	Definition	Format	Constraints
ID_SITE	Unique identifier for each site	Numeric	Positive integer
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	None
latitude	Latitude of the site, given in decimal degrees, where N is positive and S is negative	Numeric	Values between -90 and 90
longitude	Longitude of the site, given in decimal degrees, where E is positive and W is negative	Numeric	Values between -180 and 180
elevation	Elevation of the site, in meters above sea level	Numeric	None
ID_SITE_TYPE	Unique identifier of the site type (related to site_type table)	Text	Selected from predefined list
ID_BASIN_SIZE	Unique identifier of the basin size for each site (related to basin_size table)	Text	Selected from predefined list
ID_CATCH_SIZE	Unique identifier of the catch size for each site (related to catch_size table)	Text	Selected from predefined list

Table 2. Characteristics of the **site_type** table

Field label	Definition	Format	Constraints
desc_site_type	Description of the site type (e.g. marine, lacustrine-natural open water, lacustrine-volcanic lake)	Text	Selected from predefined list
ID_SITE_TYPE	Unique identifier of the site type (e.g. LTEC for lacustrine-natural open water-tectonic lake or LVOL for lacustrine-volcanic lake)	Text	Selected from predefined list
ID_HIGHER	Unique identifier of the site type in a wider classification (e.g LACU for any the lacustrine site type)	Text	Selected from predefined list

Table 3. Characteristics of the **basin_size** table

Field label	Definition	Format	Constraints
basin_desc	Description of the basin size (e.g. very small [<0.01 km2])	Text	Selected from predefined list
ID_BASIN_SIZE	Unique identifier of the basin size (e.g. VESM for a very small [0.01 km2] basin size)	Text	Selected from predefined list

Table 4. Characteristics of the **catch_size** table

Field label	**Definition**	**Format**	**Constraints**
catch_size	Description of the catch size (e.g. small [<10 km2])	Text	Selected from predefined list
ID_CATCH_SIZE	Unique identifier of the catch size (e.g SMAL for a small [<10 km2] catch size)	Text	Selected from predefined list

Table 5. Characteristics of the **entity** table

Field label	**Definition**	**Format**	**Constraints**
ID_SITE	Unique identifier for each site (as given in the site table)	Numeric	Positive integer
ID_ENTITY	Unique identifier for each entity, defined as separate record or sampling point within a site	Numeric	Positive integer
entity_name	Entity name as given by original authors or as defined by us where there was no unique name given to the entity	Text	None
latitude	Latitude of the entity, given in decimal degrees, where N is positive and S is negative	Numeric	Values between -90 and 90
longitude	Longitude of the entity, given in decimal degrees, where E is positive and W is negative	Numeric	Values between -180 and 180
elevation	Elevation of the entity, in meters above sea level	Numeric	None
ID_ENTITY_TYPE	Unique identifier of the entity type (related to entity_type table)	Text	Selected from predefined list
source	Source of the pollen data	Text	Selected from predefined list

Table 6. Characteristics of the **entity_type** table

Field label	**Definition**	**Format**	**Constraints**
entity_type	Description of the entity type (e.g. lacustrine core, moss polster or moss, pollen trap)	Text	Selected from predefined list
ID_ENTITY_TYPE	Unique identifier of the entity_type (e.g. LACO for lacustrine core, or MOSS for moss polster or moss)	Text	Selected from predefined list
mod_or_0ky_class	Unique identifier of the entity_type at higher classification	Text	Selected from predefined list
comments	Comments on the entity type	Text	Selected from predefined list

Table 7. Characteristics of the **entity_pub** table

Field label	**Definition**	**Format**	**Constraints**
ID_ENTITY	Unique identifier for each entity (as given in the entity table)	Numeric	Positive integer
ID_PUB	Unique identifier for each publication associated to the pollen sample or record (related to pub table)	Numeric	Positive integer
ID_ENTITY_PUB	Unique identifier for each entity with a publication associated	Numeric	Positive integer

Table 8. Characteristics of the **pub** table

Field label	**Definition**	**Format**	**Constraints**
citation	Complete reference to the publication associated with the sample or pollen record	Text	None
ID_PUB	Unique identifier for each reference	Text	Positive integer

Table 9. Characteristics of the **date_info** table

Field label	**Definition**	**Format**	**Constraints**
ID_DATE_INFO	Unique identifier for each dated age	Numeric	Positive integer
ID_ENTITY	Identifier for each entity (as given in the entity table)	Numeric	Positive integer
avg_depth	Average depth in the sedimentary sequence where the sample for dating was taken, given in meters	Numeric	Positive decimal
thickness	Thickness of the sample taken for dating	Numeric	Positive decimal
lab_num	Unique identifier code for each dated sample as given by the dating laboratory	Text	None
ID_MAT_DATED	Unique identifier of the dated material (related to mat_dated table)	Text	Selected from predefined list
dated_age	Uncalibrated age of the dated sample, given in years	Numeric	Positive integer
error_positive	Positive uncertainty of the uncorrected age of the dated sample, given in years	Numeric	Positive integer
error_negative	Negative uncertainty of the uncorrected age of the dated sample, given in years	Numeric	Positive integer
ID_DATE_TYPE	Unique identifier of the method used for dating (related to date_type table)	Text	Selected from predefined list
date_comments	Comments on the dated sample (e.g. contamination suspected), obtained from the publications	Text	Selected from predefined list

Table 10. Characteristics of the **mat_dated** table

Field label	**Definition**	**Format**	**Constraints**
mat_dated	Description of the dated material (e.g. plant macrofossil, foraminifera, bulk sediment-peat, bulk sediment-calcareous lake deposits)	Text	Selected from predefined list
ID_MAT_DATED	Unique identifier of the dated material (e.g. PLMA for plant macrofossil or BUPE for bulk sediment-peat)	Text	Selected from predefined list
ID_MAT_DATED_HIGH	Unique identifier of the dated material in a wider classification (e.g. BULK for any bulk sediment material)	Text	Selected from predefined list

Table 11. Characteristics of the **date_type** table

Field label	**Definition**	**Format**	**Constraints**
date_type	Description of the method used for dating (e.g. C14, annual lamination, tephra)	Text	Selected from predefined list
ID_DATE_TYPE	Unique identifier of the method used for dating (e.g. C_14 for C14, ANNL for annual lamination or TEPH for tephra)	Text	Selected from predefined list

Table 12. Characteristics of the **sample** table

Field label	**Definition**	**Format**	**Constraints**
ID_SAMPLE	Unique identifier for each pollen sample	Numeric	Positive integer
ID_ENTITTY	Unique identifier for each entity (as given in the entity table)	Numeric	Positive integer
sample_name	Unique identifier code for each pollen sample as the source from which the data were obtained	Text	None
avg_depth	Average depth of the sample in the sedimentary sequence, given in meters	Numeric	Positive decimal
ID_SAMPLE_TYPE	Unique identifier of the sample type (related to sample_type table)	Numeric	Selected from predefined list
count_type	Format in which the pollen data of each sample are given (e.g. raw count, percentages)	Text	Selected from predefined list

Table 13. Characteristics of the **sample_type** table

Field label	Definition	Format	Constraints
sample_type	Sedimentary material of the pollen sample (e.g. lacustrine clay with shell, marine sapropel)	Text	Selected forpredefined list
ID_SAMPLE_TYPE	Unique identifier for each sedimentary material (e.g. CLSH for lacustrine clay with shell, or MASA for marine sapropel)	Text	Selected forpredefined list

Table 14. Characteristics of the **pollen_data** table

Field label	Definition	Format	Constraints
ID_SAMPLE	Unique identifier for each pollen sample (as given in the sample table)	Numeric	Positive integer
taxon_clean	Taxon name	Text	None
ID_SAMPLE_TAXA	Unique identifier for each taxon name	Numeric	Positive integer
taxon_count	Pollen count of each taxon in each sample (raw count or percentage)	Numeric	Positive decimal

Table 15. Characteristics of the **age_model** table

Field label	Definition	Format	Constraints
ID_SAMPLE	Unique identifier for each pollen sample (as given in the sample table)	Numeric	
est_age_provided	Age of the sample provided by the authors	Numeric	positive decimal
est_age_original	Age of the sample obtained from the source of the pollen data	Numeric	positive decimal
est_age_bacon_intcal13	Median age of the sample, calibrated using the IntCal13 curve	Numeric	positive decimal
est_age_bacon_intcal13_max	Upper bound of the 95% confidence interval for the median age, calibrated using the IntCal13 curve	Numeric	positive decimal
est_age_bacon_intcal13_min	Lower bound of the 95% confidence interval for the median age, calibrated using the IntCal13 curve	Numeric	positive decimal
est_age_bacon_intcal20_mean	Mean age of the sample, calibrated using the IntCal20 curve	Numeric	positive integer

est_age_bacon_intcal20_median	Median age of the sample, calibrated using the IntCal20 curve	Numeric	positive integer
est_age_bacon_intcal20_uncert_5	Lower bound of the 95% confidence interval for the median age, calibrated using the IntCal20 curve	Numeric	positive integer
est_age_bacon_intcal20_uncert_95	Upper bound of the 95% confidence interval for the median age, calibrated using the IntCal20 curve	Numeric	positive integer
est_age_bacon_intcal20_uncert_25	Lower bound of the 75% confidence interval for the median age, calibrated using the IntCal20 curve	Numeric	positive integer
est_age_bacon_intcal20_uncert_75	Upper bound of the 75% confidence interval for the median age, calibrated using the IntCal20 curve	Numeric	positive integer
comment	Comments on the age of the sample	Text	none