Scientific knowledge accumulates through collecting independent clues. This process requires obtaining results from independent data and identical or comparable analytical approaches. This is often referred as repeatability, which should not be confused with reproducibility that is the replication of the published results by using original analysis programs on the original data (Cassey & Blackburn 2006). Reproducibility can allow others (readers, reviewers, editors) to verify findings, conduct alternate analysis of the same data or similar analysis of independent data. Thus, there is an increasing need for reproducibility in areas, where results are product of large comprehensive databases, complex statistical computations and have high social or policy relevance (Parr 2007, Peng et al. 2006).

In biogeography, data sets are often comprised of large scale distribution data of multiple taxa, with various background information for environmental and taxon specific variables. The analysis of these data is also often complicated, which requires consistent input formats. These input formats are rarely match the format of the raw data. So the measured raw data need to be processed to get analytic data. Then analytic data is analysed, and computational results (numerical results, tables, figures) are produced, to be included into a scientific paper.

Figure 1 shows the research pipeline from raw data to paper. Out of the outlined stages, reproducibility requires the analytical data and the analytic code to be available as a minimum standard (Peng et al. 2006). Although full reproducibility may include other elements (measured data, processing code, computational results, and presentation code) as well. Gentleman and Temple Lang (2004) described the idea of “compendium” as a way to publish reproducible research in a dynamic document. In a compendium, readers with casual interest can find the finished paper, while more interested readers can dig deeper into the data and computation.

Figure 1. The research pipeline as a model for reproducible research (after Peng et al. 2006). The author and the reader are approaching from opposite directions. Readers may want to dig deeper into the research to verify results or conduct similar study. Minimum reproducibility can be achieved by providing measured or analytic data and the analytic code. Processing code can also be necessary when original measured data are provided. Full reproducibility would include all stages of the pipeline. The ‘mefa’ R package can be useful in data processing and analysis.
There are existing initiatives for data sharing in biogeography (e.g. Global Biodiversity Information Facility, http://www.gbif.org; Biological Collection Access Services, http://www.biocase.org/), and analytical data are often presented as supplementary on-line materials. But data processing and analysis is often described in the materials section (or in the appendix) without software code that would enable full reproducibility. We are not arguing that all studies should be published in a fully reproducible manner, because there are various reasons for data holding as well (e.g. too much effort required, need to protect own ability to publish, etc.; Campbell et al. 2002). But the reproducibility is getting to become standard in areas like public health (Peng et al. 2006), bioinformatics (Gentleman & Temple Lang) and ecology (Parr & Cummins 2005), further data sharing and reproducibility are likely to increase citations and impact factor of journals, and thus indirectly the reward to individual researchers (Hajjem et al. 2005, Parr 2007).

As part of our efforts to make our research more reproducible, we developed the 'mefa' package (Sólymos 2008) for data handling in ecology and biogeography. This is an extension to the R language and environment (R Development Core Team 2008). Besides the advanced statistical capabilities of R, it also provides means for reproducible research ('Sweave' package to combine R and LaTeX code, Leisch 2002; 'odfWeave' to combine the open document format and R code, Kuhn & Weaston 2008; R extension checking and building facilities to create compendiums, Gentleman & Temple Lang 2004). The name 'mefa' refers to the term "metafaunistics" indicating our aim to provide basic object classes and helper functions to make data processing easier, because data processing can be time consuming and a critical step in the research process.

Figure 2 shows basic object classes provided by the package. The class 'sscount' is a long table with columns for identifiers of samples, species and segments (e.g. year of collection, cf. heat colors). Other objects contain count data of the 'sscount' object in cross tabulated format for single ('xcount') or multiple ('xclist') segments (cf. heat colors). When sample and species specific attribute tables are attached, the resulting objects are of class 'mefa' or 'mflist'. These can be translated into each other, and has 'print' and 'plot' methods. These objects are often project-oriented representations (subsets) of larger data sets, to be used in a specific report.

Figure 2. Schematic representations of the object classes provided by the 'mefa' R package. The 'sscount' object has columns for identifiers of samples, species and segments (e.g. year of collection, cf. heat colors). Other objects contain count data of the 'sscount' object in cross tabulated format for single ('xcount') or multiple ('xclist') segments (cf. heat colors). When sample and species specific attribute tables are attached, the resulting objects are of class 'mefa' or 'mflist'. These can be translated into each other, and has 'print' and 'plot' methods. These objects are often project-oriented representations (subsets) of larger data sets, to be used in a specific report.
geographical context, it may contain information on the collecting event (e.g. name of collector, source of the data, date of observation). We can differentiate between taxon and sample oriented approaches. Data in the taxon oriented approach are assembled by the species identifiers (e.g. occurrences of species with latitude-longitude data), and samples (locations, grids) are not necessarily comparable in terms of sampling effort. This is common in biogeography. While in the sample oriented approach, samples are often taken in a hierarchical design, and samples are directly comparable in terms of sampling effort. This latter is most often the case in ecology.

To demonstrate some of the basic functionalities of the 'mefa' package, we use here the land snail distribution database of the Hungarian Natural History Museum (HNHM), as an example of the taxon oriented approach (example on the sample oriented approach is given in the demo of the package). The 'sscount' object contained codes of the 10×10 km UTM cell, name of species, year of collection, count values were uniform to represent voucher specimens. Figure 3 shows data accumulation based on 34,113 museum records of 136 species in 602 10×10 km UTM grid cells (9,086 species-in-grid-cell records) (Fehér & Gubányi 2001). There is a steep increase in sampling effort from 1950 to 1990 (Figure 3a) resulting in significant increase in sampling coverage (Figure 3b). On the contrary, the number of land snail species known from Hungary increased only slightly due to the high sampling intensity in the latest period (Figure 3c).

By using the sample, species and segment identifiers, we can fill a three dimensional array with the count values. The sample/species cross tabulation for a single segment (or combination of multiple segments, or combining all segments) is an object of class 'xcount'. We can than attach attribute tables for the samples and species, containing sample (location) and species specific variables (subsetted and ordered according to row and column identifier of the 'xcount' cross tabulation). The bundle of an 'xcount' object and one or two attribute tables make up a 'mefa' object. For the land snail data, we used the coordinates of the UTM cell centroids as sample attributes. The 'xcount' and 'mefa' objects contain instantly the marginal statistics (number of records and occurrences per samples and species) of the cross tabulation, that can be readily used in exploration, visualisation and subsequent analysis.

Figure 3. Temporal trends in data accumulation plotted as number (blue bars) and cumulative number (red lines) of records (a), 10×10 km UTM cells covered (b), and species (c). Vertical lines represent years with quartiles of the total data volume. The plot was made from an 'sscount' object by the 'accumulate' function.
However, we might want to combine different segments by forming non-inclusive or nested partitions of the three dimensional data to compare the “slices” with each other (see heat colors in Figure 2). The result of this approach is an object of class ‘xclist’ (or ‘mflist’ if containing attribute tables). By using the data of an ‘xclist’ object, we can explore and visualize trends e.g. in data accumulation. We used year of collection to split the database into temporally subsequent nested partitions by 10% data volume increments. Figure 4 shows the results for the first 10% and the full data set. The effect of temporal data accumulation on the relationship can be viewed as animation, to see the 100-years collecting process in its dynamism (http://www.univet.hu/users/psolymos/animefa/). What we can see is the serious over-collection of the most species rich UTM grid cells (Figure 4a). While for cells with intermediate and low species richness, the increase in the number of records is more linearly proportional to the number of species. This pattern is a result of preferences of collectors towards species rich and more natural regions, and areas close to home settlements (Sólymos 2007). The range of the initial (10%) and the final richness values are very similar. When inspecting species (Figure 4b), we see a more linear relationship between the number of records and the number of occupied grid cells. Thus, based on absolute values, none of the species are over-represented and their relative ranks are highly congruent through time. But for species, the range of the initial 10% and the 100% data differ significantly.

We can also explore this spatio-temporal collecting process on a map (Figure 5). Over-representation of “good” sites can be referred as the biotic impediment, because, with time, we have disproportionately more data from areas where we have already had data. While on the opposite, unexplored areas are disappearing more slowly, which is termed as the Wallacean shortfall. This is a common phenomenon in historical distribution databases, because most of the collectings were not planned, but rather driven by available opportunities and individual preferences for regions. From the 1960s, collecting was informally coordinated by László Pintér, former curator of Molluscs at HNHM. Results of this work were published later (Pintér et al. 1979) amongst the first national assessments of the European Invertebrate Survey.
The large database is now being reorganised along with materials of other Hungarian museum collections to be used in predictive modelling. The 'mefa' package was primarily developed to ease the processing of field data of complex hierarchical sampling designs, and large scale distribution data. Reports can be generated from 'mefa' objects in plain text or pre-formatted LaTeX format. This feature can be useful e.g. to produce supporting on-line materials for manuscripts. Current stable version of the package is 1.1, available on CRAN (http://cran.r-project.org). The development version will soon include the multi-segment object classes and the animation features that are now available as patch at http://mefa.r-forge.r-project.org. At this site, the documentation, mailing lists and news can also be found. Comments on the package and possible new feature requests are especially welcome.

We believe that reproducible research in biogeography will be more commonplace in the near future. Most of the means by which data sharing and reproducibility can be achieved are available (e.g. http://thedata.org and http://ecoinformatics.org). But there is also a need for the development of specific data manipulation techniques to be used in biogeography, for which the R software environment is ideal because it is highly extensible, free and platform independent.

References


If you want to comment on this article go to http://biogeography.blogspot.com/2008/02/recent-progress-toward-understanding.html