https://doi.org/10.22498/pages.25.2.96

The big data revolution and paleoecology

Sandy P. Harrison

Big data has revolutionized science. Cultural and practical issues have limited its impact on paleoecology, despite the field's long history of data synthesis. We need stakeholder interactions and outside-the-box thinking to maximize scientific benefits in the big data era.

Access to increasingly large quantities of data and enhanced data sharing through open-access databases have revolutionized many areas of science. The huge volume of astronomical observations generated by the Gaia mission have contributed to advances in fundamental physics. The explosion of human genomics data has led to better understanding of the causes of diseases and the development of personalized treatments. Multi-sensor Earth-observation data are being used to understand climate variability better and monitor environmental responses to changes in atmospheric composition, land use and climate. Connecting climate observations with economic data is enabling the implementation of sustainable agricultural practices; connecting climate information with energy-sector data is allowing projections of the response to climate variability to be factored into energy management practice. Thus, the big data revolution is not just about the amount of data or the use of high-powered statistics. It is about data being exploited to answer completely different types of questions from the ones for which they were originally collected.

Paleoenvironmental datasets have a long history, starting with the datasets created by CLIMAP (Climate: Long range Investigation, Mapping and Prediction) and COHMAP (Co-operative Holocene Mapping Project) in the 1970s and 80s. Several community databases originated in the 1980s, including the Global Lake Status Database (Street-Perrott et al. 1989), the International Tree Ring Database (https://data.noaa.gov/dataset/international-tree-ring-data-bank-itrdb) and the European (www.europeanpollendatabase.net) and North American (www.ncdc.noaa.gov/data-access/paleoclimatology-data/datasets/pollen) Pollen Databases. Archives have subsequently been created for other kinds of paleoenvironmental records. These databases facilitate comparisons among records, regional paleoecological and paleoclimatic reconstructions, evaluation of paleoclimate modeling results and other applications.

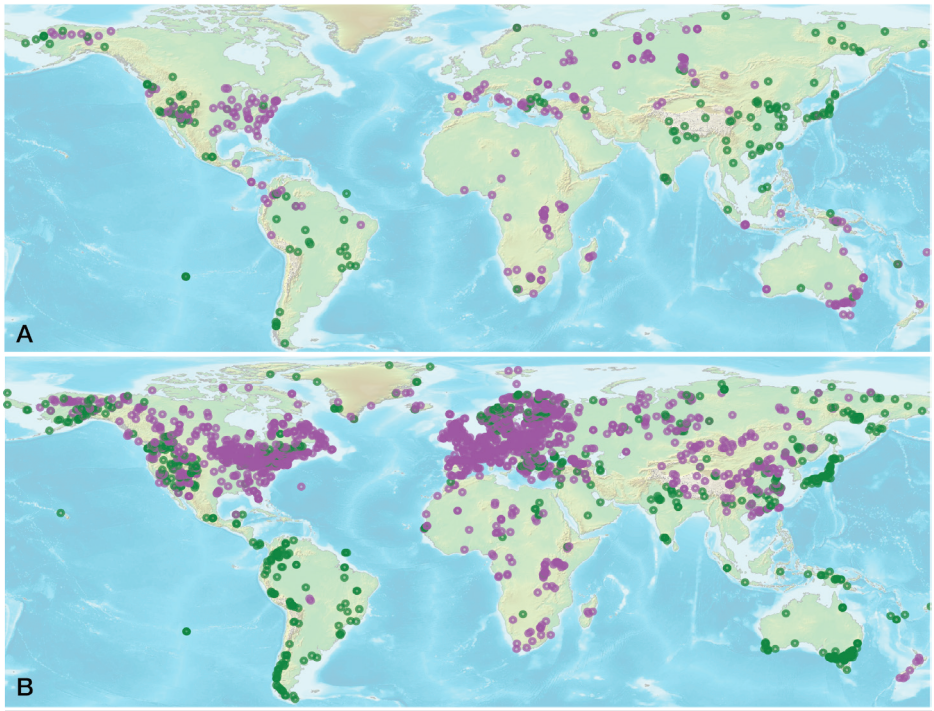

Figure 1: Pollen data are the most widely-distributed source of quantitative paleoclimate reconstructions used to evaluate model simulations of the mid-Holocene (MH) and Last Glacial Maximum (LGM). However, there are large gaps in the data coverage although many more pollen sequences are available from public-access databases that could be used for reconstructions. The maps show the distribution of sites for **(A)** LGM and **(B)** MH, where magenta dots represent sites with climate reconstructions (Bartlein et al. 2011; Prentice et al. 2017), and green dots represent pollen sites where it would be possible to make quantitative reconstructions (data from BIOME 6000 database; https://doi.org/10.17864/1947.99 and from the EMBSeCBIO database; Cordova et al. 2009).

Benefits of (big) data sharing

Data sharing is now firmly embedded in the scientific culture. Several factors have contributed to this development, including the activities of the research groups developing databases, national and international funding agency policies that mandate open-access publication and data archiving, journal rules that increasingly specify that original data must be available for scrutiny and replication of results, the increasing number of journals dedicated solely to publishing data sets, the increasing ease of obtaining persistent identifiers for data sets, and the recognition that openness about data increases research impact (Piwowar and Vision 2013). However, these advances have not led to a revolution in the approach to data in paleoecology. There are regional analyses (e.g. Huntley et al. 2013) and some global analyses of paleoecological data sets (e.g. Daniau et al. 2012). But the scientific focus is still largely on documenting changing vegetation patterns (e.g. Prentice et al. 2000) or reconstructing climate (e.g. Bartlein et al. 2011) – goals that date from the 1970s and provided the original motivation for the construction of paleoecological databases. Far more could be done using the data that now exist (Fig. 1).

As one example, a large community of ecologists focuses on plant functional traits and how community-mean trait values change along environmental gradients. At a fundamental level, this research seeks to explain aspects of the function of plants and ecosystems (Ali et al. 2015). Trait-based analyses are also being used to explore the controls on within- and between-site diversity (Ackerly and Cornwell 2007) and to develop vegetation models based on fundamental principles rather than empirical relationships (Fyllas et al. 2014). Ecologists are also exploring how species and ecosystems respond to climate change, on shorter (acclimation) and longer (adaptation, migration) timescales. There is growing literature on how the velocity of climate change affects species' potential for adaptation and migration (e.g. Loarie et al. 2009) and extinction risks for different groups of organisms (Settele et al. 2014). These are all important questions, with implications for conservation policy; paleoecological data should have a great deal to say about all of them.

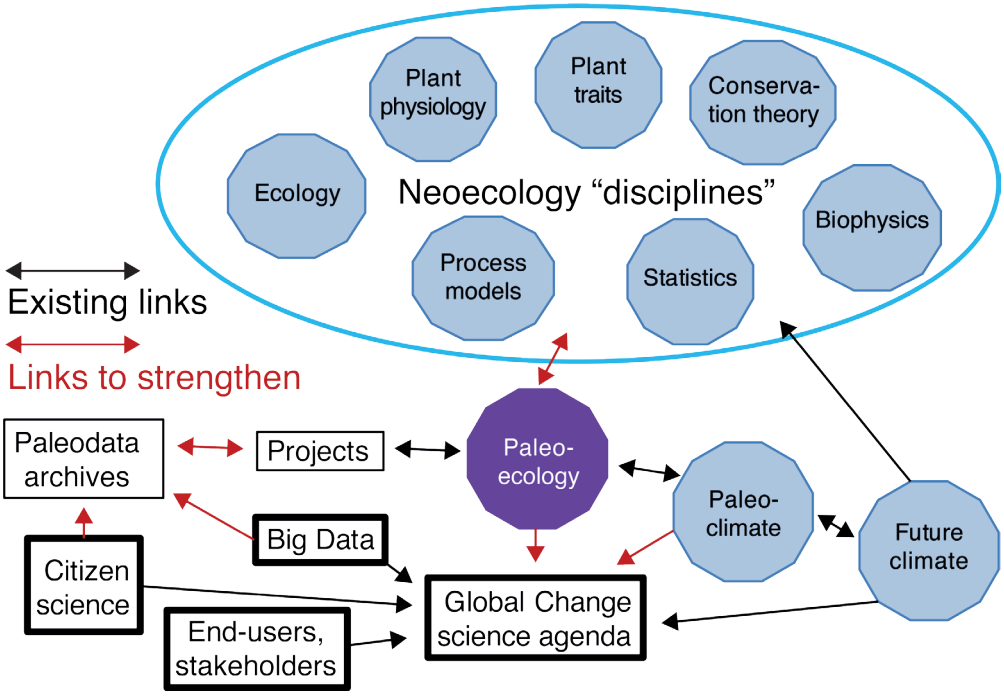

Figure 2: Conceptual diagram of the current and potential position of paleoecology (purple polygon) in the global change scientific framework, showing areas where relationships to other sciences (blue polygons), data sources and stakeholders could be strengthened to optimize the value of paleoecology to address real-world issues.

What stops us embracing big data?

Why has paleoecology missed out on the big data revolution? Contributory factors include the labor-intensive nature of data generation, lack of specialized training in data analysis and modeling techniques, and a persistent lack of cross-fertilization with contemporary ecology. Paleoecology has a strong site-based focus, and a tendency for practitioners to specialize in a particular group of organisms and a particular study region. This is understandable to some extent: the faunas and floras of each continent are different; hard-won expertise in the identification of one group of subfossil organisms does not help with other groups. However, in ecology generally, theoretical data-analysis and modeling approaches are well-established fields of endeavor. Paleoecology, by contrast, is still largely a field- and laboratory-based science; scientists are expected to serve an apprenticeship that involves primary data collection but generally does not provide training in the quantitative and data-analytical skills necessary to make sense of large data sets.

Data-generating techniques in paleoecology are notoriously time-consuming and this reinforces the site-based focus as well as limiting the amount of data that exists. Contemporary ecology and ecosystem science are benefiting from massive new data sources involving automated retrieval – from drones to satellites. Automation in paleoecology, for example in pollen counting, has been discussed repeatedly but there has been little concrete progress. Ecologists have also harnessed the power of citizen science to generate large data sets with high temporal resolution. Activities such as Climateprediction.net (www.climateprediction.net) and Zooniverse (www.zooniverse.org) show it is possible to involve non-specialists in scientific projects and generate valuable data on a scale otherwise impossible. We need to think creatively about harnessing people's enthusiasm for science.

The analysis of large data sets requires skills including working with database software, advanced statistical methods and multivariate analysis. Training in these skills at undergraduate and postgraduate level is patchy. Furthermore, the growing importance of quantitative models as a means to embody and test hypotheses puts a premium on mathematical and programing competencies that are increasingly prioritized in the training of ecologists and evolutionary biologists, but generally not in the training of paleoecologists.

A future for paleoecology

"The present is the key to the past; the past is the key to the future". Everyone says it, but how often does this crossover occur? There is very little interaction between paleoecology and developments in contemporary ecology, ecophysiology and biophysics. The Future Earth program (http://futureearth.org) could provide opportunities to embed paleoecology more firmly in a multidisciplinary Earth system science context (Fig. 2). Future Earth's stated commitment to "involving stakeholders throughout the entire research process from co-design to dissemination" should provide opportunities for scientists with different backgrounds, including the unique temporal perspective on species and ecosystems which paleoecologists provide, to work together towards the solution of real-world problems arising from global environmental change. But the realization of these aspirations will require paleoecologists, and others, to think "outside the box" and pay attention to other disciplines.

If paleoecology is to survive, we need a revolution in our definition of the legitimate sphere of investigation and our approaches to training the next generation of paleoecologists. We need to think creatively about generating paleoecological data efficiently and also about the questions that can be addressed with paleoecological data. We need to talk to scientists and practitioners from related fields to co-design research that will realize the unique contribution that paleoecological observations could make to Earth system science and management.

AFFILIATIONS

Centre for Past Climate Change and School of Archaeology, Geography and Environmental Sciences (SAGES), Reading University, UK

CONTACT

Sandy P. Harrison: s.p.harrison@reading.ac.uk

REFERENCES

Ackerly DD, Cornwell WK (2007) Ecol Lett 10: 135-145

Ali AA et al. (2015) Ecol Appl 25: 2349-2365

Bartlein PJ et al. (2011) Clim Dyn 37: 775-802

Cordova CE et al. (2009) Quat Int 197: 12-26

Daniau A-L et al. (2012) Glob Biogeochem Cycl 26: GB4007

Fyllas NM et al. (2014) Geosci Model Dev 7: 1251-1269

Huntley B et al. (2013) Quat Sci Rev 70: 158-175

Loarie SR et al. (2009) Nature 462: 1052-1055

Piwowar HA, Vision TJ (2013) PeerJ 1: e175

Prentice IC et al. (2000) J Biogeog 27: 507-519

Prentice IC et al. (2017) Glob Planet Change 149: 166-176

Settele J et al. (2014) Terrestrial and inland water systems. In: Field CB et al. (Eds) Climate Change 2014: Impacts, Adaptation, and Vulnerability. Cambridge University Press, 271-359

Street-Perrott FA et al. (1989) U.S. DOE/ER/60304-H1 TR046. US DOE Tech Rep, 213pp