Journal of Vegetation Science **27** (2016) 865–867

COMMENTARY

Vegetation science in the age of big data

Scott L. Collins

Collins, S.L. (corresponding author, scollins@sevilleta.unm.edu)

Department of Biology, MSC03-2020, University of New Mexico, Albuquerque, NM, 87131, USA

Abstract

The age of big data is poised to revolutionize vegetation science. As online resources continue to grow, vegetation ecologists will need a growing set of computational skills to advance vegetation science in the digital age. Two papers in this issue of the *Journal of Vegetation Science* (Wiser 2016, Sandel et al. 2016) illustrate the resources available and use of big data to explore challenging ecological questions.

The science of ecology has gone through several broadly defined developmental phases over the past few centuries. Starting from the age of exploration led by luminaries such as von Humboldt and Darwin, plant ecology moved through a period during much of the 20th century that could be characterized by predominantly descriptive, observational field studies. Only relatively recently have ecologists truly embraced the widespread use of field experiments to test hypotheses under some degree of realism. Regardless of approach, the development and testing of theory, either through observations or experiments, have been at the core of ecology since its inception. Now that the information age is upon us, ecology is going retro as we enter a new age of exploration, only this time it is digital and can easily encompass long time scales and large spatial scales.

For much of its rich history, research teams in ecology were generally small and often geographically isolated. Although the International Biological Program ushered in the age of 'Big Ecology' for ecosystem science (Coleman 2010), most population and community ecologists operated in very small teams. Indeed, the Web of Science lists 32 publications by the revered plant ecologist J.T. Curtis. Of these, only four papers had more than two authors (three, actually), and much of his work centered around the vegetation of Wisconsin. The story is much the same for the eminent ecologist R.H. Whittaker, who was highly collaborative for his time, yet Whittaker had only eight papers with three or more authors out of 67 listed in the database, and four of those were on Hubbard Brook. Now multiple authored papers on global scale questions in ecology are more and more common (Nabout et al. 2015) as a consequence of vast amounts of on-line data probed by ecologists with sophisticated programming skills, coupled with a significant change in culture that now values synthesis, collaboration and data sharing, all of which are facilitated via global connectivity.

Two papers in this issue reflect different aspects of this 'new normal' in ecology. Wiser (2016) describes in wonderful detail the Herculean efforts needed not only to build and connect a growing number of vegetation-plot databases, but also the monumental challenges associated with downloading and especially integrating data from multiple data sources. A second paper in this issue (Sandel et al. 2016) on the global trait distribution and trait relationships of grasses provides an excellent example of how the variety of information available in databases can be mined and combined for data exploration that advances big picture understanding in vegetation science.

Creating a user-friendly database is a huge challenge. Wiser (2016) identifies an amazing 231 databases with more than three million plot records in the Global Index of Vegetation-Plot Databases. She describes the on-going efforts to build and integrate these resources, as well as development of tools that facilitate wrangling the data into analysable formats and that help to standardize nomenclature across data sets. Although you would think that the basic format of vegetation data would be relatively easy to standardize, data are instead stored in a variety of formats, and most databases were created with specific and differing goals in mind, leading to considerable heterogeneity in data formats. More importantly, creating a database is not an end unto itself. A database exists for more than just managing and storing data, but instead, the goal should be to make data both discoverable and *usable* hopefully in perpetuity.

Information Management has become its own discipline to deal with the myriad challenges of documenting, describing, storing, protecting, managing and versioning data, plus making it discoverable. Yet as we have learned over the past 25 yrs, more and more sophisticated

Doi: 10.1111/jvs.12459

Informatics tools do not necessarily mean that data are usable. Often data are very raw and in unique formats in different data sources. Thus, many different software tools are needed to manipulate raw data drawn from multiple sources to make them compatible and analysable. Rarely are these tools linked to the data themselves nor are they made freely available because they are often developed for one-off analyses. This needs to change, and the open science movement is pushing ecology in that direction (e.g. Hampton et al. 2015; Mislan et al. 2016).

Many challenges remain. Metadata standards may not be sufficient to adequately describe the data. Investigators that contribute data to a repository may place restrictions on their use. Others are simply resistant to open sharing of data (Mills et al. 2015). For this situation to improve, the community needs to rally around explicit metadata standards, such as Ecological Metadata Language, to facilitate data discovery and use. Advances in the ethics of data sharing and data use need to be developed and credible mechanisms for assigning attribution and credit are needed. Citable data sets with Digital Object Identifiers (DOIs) are emerging as one solution but this will not satisfy those who consider their data to be their own intellectual property. We have a long way to go, but Wiser's thoughtful paper should be read widely. My favourite line comes near the end, 'Archiving data is especially important for those at later career stages to ensure that their legacy endures.' I'm not convinced that senior citizens are the only constituency holding a vast array of dark data. In truth, this message is important at all career stages so that archiving and sharing becomes a habit for all ecologists. Overall, Wiser's presentation should elevate the discussion and motivate solutions for the use of big data in vegetation science.

Sandel et al. (2016) provide an excellent example of the use of big data in vegetation science. To conduct their analysis, Sandel et al. (2016) combined plant distribution and trait data from GrassBase, USDA Plants Database, VegBank, the Kew Gardens Seed Information Database, the C3/C4 database and the global plant trait database TRY, along with additional information they extracted from the primary literature. This is no simple task as described by Wiser (2016). Sandel et al. (2016) used the information housed in these databases to explore the relationships among 14 mostly quantitative functional traits of grasses, and how these relationships 'translate to trait–trait correlations of abundance-weighted means of assemblages.' By varying spatial grain they then assessed how well trait–trait correlations of grasses at the assemblage level reflected correlations at the species level.

Sandel et al. (2016) find two clear clusters of traits in the grasses. The first cluster relates to plant size and is defined by leaf size, seed mass, plant height and rooting depth. The second cluster echoes the leaf economics spectrum with somewhat weaker correlations between SLA and leaf N and P concentrations. In general, these relationships held at the assemblage and species levels of resolution, suggesting that assemblage-level patterns are constrained by species-level trait correlations.

It is indeed comforting to know that the traits of the world's grasses conform to the general spectrum of plant trait distributions (Díaz et al. 2016). And yet, much more needs to be done. For one thing, the analysis of Díaz et al. (2016) used only above-ground traits, a limitation that also affected the analysis of Sandel et al. (2016). This limitation exists because it is very difficult to get comparable below-ground trait data, and yet it is fundamentally important for plants, in general, and grasses specifically because in many ecosystems there is more grass biomass below-ground than above-ground. This suggests that for some extensive ecosystems, at least, much of the action in grasses is in the soil and less so in the air.

Second, the data set of Sandel et al. (2016) could be more fully explored with a more ecological context in mind, such as fire and grazing. For example, Forrestel et al. (2014, 2015) examined the taxonomic, phylogenetic and functional responses of grasses in North American and South African savanna grasslands to changes in fire and grazing regimes. They found phylogenetic clustering in both continents under high fire frequencies, driven primarily by species in the Andropogoneae, and a narrow range of functional strategies related to post-fire regeneration. Similarly, they found that functional syndromes associated with grazing resistance were conserved in both sites. They conclude that grazing and aridity act together as selective forces on grass functional traits.

As the age of big data in vegetation science continues, many challenging questions remain to be explored. One of the earliest conceptual battles in North American community ecology was the Clements–Gleason debate about the nature and structure of 'climax' plant communities, a debate of such prominence that it still appears in general biology textbooks. The caricatures used to illustrate the two competing theories (e.g. Collins et al. 1993) are modern and perhaps extreme interpretations of verbal models. Neither Clements nor Gleason illustrated their models graphically. Of course the debate has been settled and most vegetation scientists are Gleasonian. But accumulating evidence against the Clementsian community-unit model does not actually quantitatively define the nature and structure of the continuum. Much can be learned by bringing new tools to bear on an old question, and now we have in place the analytical tools and the big data needed to quantitatively describe the 'true' nature and structure of the continuum. Doing so would

Doi: 10.1111/jvs.12459

have significant implications for understanding and predicting how plant communities will respond to climate change.

There are many important questions that are waiting to be addressed in the new age of digital exploration. This represents an exciting time where science will advance as more and more data and analytical tools come on line in an era of synthesis and analysis facilitated by extensive collaborations. In doing so, new approaches that move us away from a strict adherence to the hypothetico-deductive approach will greatly expand the types of questions that can be asked and ultimately advance theory and understanding in vegetation science. The papers of Wiser (2016) and Sandel et al. (2016) are exciting steps in this direction.

References

Coleman, D.C. 2010. *Big ecology: the emergence of ecosystem science*. University of California Press, Oakland, CA, US.

Collins, S.L., Glenn, S.M. & Roberts, D.W. 1993. The hierarchical continuum concept. *Journal of Vegetation Science* 4: 149–156.

Díaz, S., Kattge, J., Cornelissen, J.H.C., Wright, I.J., Lavorel, S., Dray, S., Reu, B., Kleyer, M., Wirth, C. (...) & Gorné, L.D. 2016. The global spectrum of plant form and function. *Nature* 529: 167–173.

Forrestel, E.J., Donoghue, M.J. & Smith, M.D. 2014. Convergent phylogenetic and functional responses to altered fire regimes in mesic savanna grasslands of North America and South Africa. *New Phytologist* 203: 1000–1011.

Forrestel, E.J., Donoghue, M.J. & Smith, M.D. 2015. Functional differences between dominant grasses drive divergent responses to large herbivore loss in mesic savanna grasslands of North America and South Africa. *Journal of Ecology* 103: 714–724.

Hampton, S.E., Anderson, S.S., Bagby, S.C., Gries, C., Han, X., Hart, E.M., Jones, M.B., Lenhardt, W.C., MacDonald, A. (...) & Zimmerman, N. 2015. The Tao of open science for ecology. *Ecosphere* 6: 1–13.

Mills, J.A., Teplitsky, C., Arroyo, B., Charmantier, A., Becker, P.H., Birkhead, T.R., Bize, P., Blumstein, D.T., Bonenfant, C. (...) & Zedrosser, A. 2015. Archiving primary data: solutions for long-term studies. *Trends in Ecology & Evolution* 30: 581–589.

Mislan, A.S., Heer, J.M. & White, E.P. 2016. Elevating the status of code in ecology. *Trends in Ecology & Evolution* 31: 4–7.

Nabout, J.C., Parreira, N.R., Teresa, F.B., Carneiro, F.M., da Chuha, H.F., de Souza Ondei, L., Caramori, S.S. & Soares, T.N. 2015. Publish (in a group) or perish (alone): the trend from single- to multi-authorship in biological papers. *Scientometrics* 102: 357–364.

Sandel, B., Monnet, A.-C. & Vorontsova, M. 2016. Multidimensional structure of grass functional traits among species and assemblages. *Journal of Vegetation Science* 27: 1047–1060.

Wiser, S.K. 2016. Achievements and challenges in the integration, reuse and synthesis of vegetation plot data. *Journal of Vegetation Science* 27: 868–879.

Doi: 10.1111/jvs.12459