perspectives in biogeography

Hypothesis testing, curve fitting, and data mining in macroecology

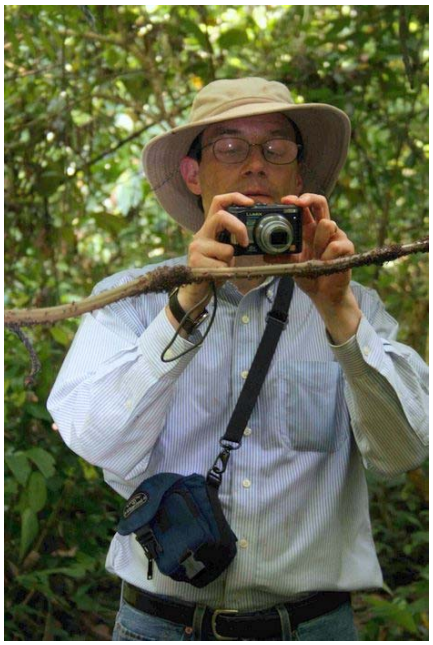
Nicholas J. Gotelli, *Department of Biology, University of Vermont, Burlington, VT 054505.* Nicholas.Gotelli@uvm.edu

Changes in technology and methodology can have a big influence on how we do science. In this essay, I will discuss how new methods for the acquisition and analysis of data have affected biogeography and macroecology.

The underlying data used by macroecologists are geo-referenced specimen collections (GBIF 2008). For many decades, biogeographers explored the globe to collect and catalog these kinds of data (e.g. Darlington 1957). The numbers, usually counts of species or maps of geographic ranges, were plotted in simple graphs and used in support of narrative explanations and historical accounts of the patterns. Explicit hypothesis-testing was rare, although pioneering analyses of taxonomic diversity indices by C.B. Williams and other European ecologists (Järvinen 1982) foreshadowed the statistical perspective that would begin to dominate ecology and biogeography in the 1970s (Gotelli and Graves 1996).

Today, the widespread availability of compiled data sets on the internet means that young scientists can begin successful careers in macroecology without ever going in the field to collect data themselves. Of course, since most of the earth's biota has not even been described taxonomically (May 1995) – much less mapped biogeographically – there is still a great deal of primary data collecting to do. But even some of this activity may become automated, with the most promising avenue being the mapping of vegetation through the use of remote sensing and satellite imagery (Gillespie et al. 2008).

With less emphasis on data collection, more energy has gone into statistical analysis and interpretation. Sophisticated methods such as spatial regression analysis (Lichstein et al. 2002) have been used to compare patterns in multiple data sets and address long-standing



hypotheses about the origin and maintenance of the latitudinal gradient in species richness (Rohde 1992, Willig et al. 2003). An entire subdiscipline of bioclimatic niche modeling has emerged as macroecologists have used species occurrence data to predict how biotas will respond to global climate change (Elith et al. 2006).

In spite of this statistical sophistication, macroecologists still have not achieved a satisfactory understanding of global patterns of species diversity (Currie et al. 2004), nor have they developed trustworthy tools for forecasting future biotic change (Araújo and Rahbek 2006). In fact, the published conclusions still sound an awful lot like the narratives of the early biogeographers! But instead of making these arguments on the basis of simple species richness plots, macroecologists make them on the size of the *p*-values or the correlation coefficients from their regression models.

There are two related problems here, one with the hypotheses and the other with the statistical methods. For the most part, hypotheses in

perspectives in biogeography

Hypothesis testing, curve fitting, and data mining in macroecology

macroecology are just verbal descriptions of mechanisms ("higher productivity in the tropics allows for more biodiversity"). But since multiple explanations can generate the same qualitative patterns ("greater temperature stability in the tropics allows for more biodiversity"), we are not going to easily distinguish these mechanisms through qualitative assessment of correlations alone.

In this regard, I think the most important recent breakthrough in macroecology has been the development of metabolic theory (Allen et al. 2002). This theory, derived from first principles that do not depend in a circular way on existing data, predicts a quantitative relationship between temperature and biodiversity. Instead of just testing a null hypothesis of a slope of zero, we can now test whether observed slopes (with appropriate transformations) deviate from -0.65, the predicted value from the model (Hawkins et al. 2007). Controversy over the empirical support for metabolic theory (Hawkins et al. 2007, Gillooly and Allen 2007) should not obscure its importance: metabolic theory makes quantitative, not just qualitative, predictions and that is what we need right now in macroecology.

Theoreticians should step up to the plate and develop quantitative theories for other hypotheses in macroecology. As recently proposed by O'Brien (2006), the water-energy model may provide an emerging framework that will generate functional forms for water and energy variables derived from first principles of physiology and physical constraints imposed by the energetics of liquid water. For now, however, these models are either entirely verbal (Vetaas 2006), or they are derived from fitted regression functions that are specific to particular taxa, spatial scales, and continents (O'Brien 1998).

In addition to the development of new theory, we need to move beyond analytical methods that simply fit curves to data and test patterns against simple statistical null hypotheses. Some macroecologists are beginning to develop stochastic simulation models that include explicit algorithms for the origin, spread, and extinction of species in a bounded geographic domain (e.g. Storch et al. 2006, Rahbek et al. 2007, Rangel et al. 2007) These mechanistic simulation models (Grimm et al. 2005) have their roots in the mid-domain effect (Colwell and Lees 2000), a pleasingly simple explanation for species richness gradients that emerged from the random placement of contiguous species ranges in a bounded domain. This kind of modeling exercise raises its own challenges: how do we empirically estimate model parameters, and how do we explore the behavior of such a model over a potentially very large parameter space? But this simulation approach may allow macroecology to move beyond statistical correlations, and can serve as a nice complement to theoretical investigations. Simulation models may even provide quantitative predictions in cases where the mathematical models do not have a tractable analytic solution.

In a provocative essay in *Wired* magazine, Anderson (2008) speculates that one day traditional hypothesis testing will be unnecessary. Some data-mining enthusiasts believe that, with enough data, correlations will reveal mechanisms in comprehensive statistical models that encompass all possible data. I think the data miners are probably right. Exciting new work in computer science has led to very sophisticated "reverse-engineering" algorithms that have great promise for uncovering the functional form of relationships among correlated variables. These new iterative methods use data partitioning, automated probing, and snipping to sequentially modify and test underlying nonlinear functions with data-rich time series.

For example, Bongaard and Lipson (2007) successfully recovered the functional form of the movement of a pendulum using as input

the temporal series of spatial coordinates of a swinging pendulum. Their algorithm repeatedly "sampled" the data set from the most critical regions (where the pendulum was changing direction) and iteratively arrived successfully at the correct equations for motion.

Interestingly, the same methods were not so successful when applied to the famous ecological time series of snowshoe hare and Canadian lynx populations (Elton and Nicholson 1942). The algorithm did generate a pair of coupled differential equations (Bongaard and Lipson 2007). However, we know that the hare-lynx cycle is not caused entirely by coupled predator-prey interactions.

The problem, of course, is not the algorithm, but the limited data that it was fed. The time series of pelt records from the Hudson Bay Company does not reveal the critical observations of hare populations on islands in eastern Canada that cycle in the absence of the lynx (Keith 1963). The analysis also did not include time series on the secondary plant compounds in tundra vegetation, which accumulate under intense grazing and may be ultimately responsible for endogenous cycles of the hare (Keith 1983). And the model did not include time-series on snowpack depth or solar sunspot activity, both of which probably contribute to the regional synchrony of hare lynx cycles (Sinclair et al. 1993).

Without such "expert knowledge" it is easy to understand why the model failed. If those data inputs were provided, I think it is very likely the model would reveal the correct functional form of the relationships among hare, lynx, vegetation, and climate. But for now, the use of passive machine-learning algorithms applied to large data sets is an inefficient way to test hypotheses and make progress in macroecology. And given the pressing need to understand how biotas will respond to climate change, I am not sure we have the luxury of waiting for these comprehensive data sets to accumulate.

Nevertheless, the paradigm of machine learning seems to be the direction that much of the bioclimatic niche modeling research is going. If the goal of this research is to understand how biotas will shift in response to climate change, I think it is going to be much more fruitful if we combine it with an experimental approach. Experimental translocation of individuals beyond their current range boundaries (Hellmann et al. 2008) and experimental manipulations of abiotic variables to mimic effects of climate change on populations and communities (Harte and Shaw 1995, Suttle et al. 1997) are very powerful approaches. Experiments can provide realistic parameter estimates for bioclimatic niche models. Even simple models that are supported by experimental data will probably be more trustworthy than sophisticated models that are not.

In sum, the availability of large data bases, the emergence of quantitative predictive theories, and the development of new computational tools and simulation methods make this an exciting time to be studying macroecology. There are pressing applied problems of global climate change that we can address with these new tools and data. And along the way, perhaps we will even answer some unresolved questions in biogeography about species richness gradients.

Acknowledgements

This essay was inspired by the work of the Synthetic Macroecological Models of Species Diversity Working Group supported by the National Center for Ecological Analysis and Synthesis, a Center funded by NSF (Grant #DEB-0553768), the University of California, Santa Barbara, and the State of California.

perspectives in biogeography

Hypothesis testing, curve fitting, and data mining in macroecology

References

Allen, A.P., J. H. Brown, J. F. Gillooly. 2002. Global biodiversity, biochemical kinetics, and the energetic-equivalence rule. Science 297: 1545-1548.

Anderson, C. 2008. The end of theory: the data deluge makes the scientific method obsolete. Wired Magazine 16.07. http://www.wired.com/science/discoveries/magazine/16-07/pb_theory

Araújo, M. B., and C. Rahbek. 2006. How does climate change affect biodiversity? Science 313: 1396-1397.

Brown, J.H., J. F. Gillooly, A. P. Allen, V.M. Savage, and G. B. West. 2004 Toward a metabolic theory of ecology. Ecology 85: 1771-1779.

Colwell, R.K., and D. C. Lees. 2000. The mid-domain effect: geometric constraints on the geography of species richness. Trends in Ecology & Evolution 15:70–76.

Currie, D. J., G. G. Mittelbach, H. V. Cornell, R. Field, J. F. Guegan, B. A. Hawkins, D. M. Kaufman, J. T. Kerr, T. Oberdorff, E. O'Brien, and J. R. G. Turner. 2004. Predictions and tests of climate-based hypotheses of broad-scale variation in taxonomic richness. Ecology Letters 7:1121-1134.

Darlington, P.J. Jr. 1957. Zoogeography: The Geographical Distribution of Animals. John Wiley & Sons, Inc.

Elith, J., C. H. Graham, R. P. Anderson, M. Dudik, S. Ferrier, A. Guisan, R. J. Hijmans, F. Huettmann, J. R. Leathwick, A. Lehmann, J. Li, L. G. Lohmann, B. A. Loiselle, G. Manion, C. Moritz, M. Nakamura, Y. Nakazawa, J. M. Overton, A. T. Peterson, S. J. Phillips, K. Richardson, R. Scachetti-Pereira, R. E. Schapire, J. Soberon, S. Williams, M. S. Wisz, and N. E. Zimmermann. 2006. Novel methods improve prediction of species' distributions from occurrence data. Ecography 29:129-151.

Elton, C. and M. Nicholson. 1942. The ten-year cycle in numbers of the lynx in Canada. Journal of Animal Ecology 11: 215-244.

Gillespie, T. W., G. M. Foody, D. Rocchini, A. P. Giorgi, and S. Saatchi. 2008. Measuring and modelling biodiversity from space. Progress in Physical Geography 32:203-221.

Gillooly, J. F. and A. P. Allen. 2007. Linking global patterns in biodiversity to evolutionary dynamics using metabolic theory. Ecology 88:1890-1894.

GBIF, Global Biodiversity Information Facility. 2008. http://www.gbif.org/press/factsheet

Gotelli, N.J. and G.R. Graves. 1996. Null Models in Ecology. Smithsonian Institution Press, Washington, DC.

Grimm, V., E. Revilla, U. Berger, F. Jeltsch, W. M. Mooij, S. F. Railsback, H. H. Thulke, J. Weiner, T. Wiegand, and D. L. DeAngelis. 2005. Pattern-oriented modeling of agent-based complex systems: Lessons from ecology. Science 310:987-991.

Harte, J. and R. Shaw. 1995. Shifting dominance within a montane vegetation community- results of a climate-warming experiment. Science 267: 876-880.

Hawkins, B. A., F. S. Albuquerque, M. B. Araujo, J. Beck, L. M. Bini, F. J. Cabrero-Sanudo, I. Castro-Parga, J. A. F. Diniz, D. Ferrer-Castan, R. Field, J. F. Gomez, J. Hortal, J. T. Kerr, I. J. Kitching, J. L. Leon-Cortes, J. M. Lobo, D. Montoya, J. C. Moreno, M. A. Olalla-Tarraga, J. G. Pausas, H. Qian, C. Rahbek, M. A. Rodriguez, N. J. Sanders, and P. Williams. 2007. A global evaluation of metabolic theory as an explanation for terrestrial species richness gradients. Ecology 88:1877-1888.

Hellmann, J. J., S. L. Pelini, K. M. Prior, and J. D. K. Dzurisin. 2008. The response of two butterfly species to climatic variation at the edge of their range and the implications for poleward range shifts. Oecologia 157:583-592.

Järvinen, O. 1982 Species-to-genus ratios in biogeography: a historical note. Journal of Biogeography 9: 363–370.

Keith, L.B. 1963. Wildlife's Ten Year Cycle. University of Wisconsin Press, Madison.

Keith, L.B. 1983. Role of food in hare population cycles. Oikos 40: 385-395.

Lichstein, J. W., T. R. Simons, S. A. Shriner, and K. E. Franzreb. 2002. Spatial autocorrelation and autoregressive models in ecology. Ecological Monographs 72:445-463.

May, R. M. 1988. How many species are there on earth?. Science 241:1441-1449.

O'Brien, E.M. 1998. Water energy dynamics, climate and prediction of woody plant species richness: an Interim General Model. Journal of Biogeography, 25, 379-398.

O'Brien, E. M. 2006. Biological relativity to water-energy dynamics. Journal of Biogeography 33: 1868-1888.

Rahbek, C., N. J. Gotelli, R. K. Colwell, G. L. Entsminger, T. Rangel, and G. R. Graves. 2007. Predicting continental-scale patterns of bird species richness with spatially explicit models. Proceedings of the Royal Society B-Biological Sciences 274:165-174.

perspectives in biogeography

Hypothesis testing, curve fitting, and data mining in macroecology

Rangel, T.F.L.V.B., Diniz-Filho, J.A.F., & Colwell, R.K. 2007. Species richness and evolutionary niche dynamics: a spatial pattern-oriented simulation experiment. American Naturalist 170: 602-616.

Rohde, K. 1992. Latitudinal gradients in species diversity: the search for the primary cause. Oikos 65:514-527.

Sinclair, A.R.E., J.M. Gosline, G. Holdsworth, C.J. Krebs, S. Boutin, J.N.M. Smith, R. Boonstra, and M. Dale. 1993. Can the solar cycle and climate synchronize the snowshoe hare cycle in Canada? Evidence from tree rings and ice cores. The American Naturalist 141: 173-198.

Storch, D., R. G. Davies, S. Zajicek, C. D. L. Orme, V. Olson, G. H. Thomas, T. S. Ding, P. C. Rasmussen, R. S. Ridgely, P. M. Bennett, T. M. Blackburn, I. P. F. Owens, and K. J. Gaston. 2006. Energy, range dynamics and global species richness patterns: reconciling mid-domain effects and environmental determinants of avian diversity. Ecology Letters 9:1308-1320.

Suttle, K. B., M. A. Thomsen, and M. E. Power. 2007. Species interactions reverse grassland responses to changing climate. Science 315:640-642.

Vetaas, O. 2006. Biological relativity to water-energy dynamics: a potentially unifying theory? Journal of Biogeography, 33, 1866-1867.

Willig, M. R., D. M. Kaufman, and R. D. Stevens. 2003. Latitudinal gradients of biodiversity: Pattern, process, scale, and synthesis. Annual Review of Ecology Evolution and Systematics 34:273-309.

If you want to comment on this article go to http://biogeography.blogspot.com/2008/10/hypothesis-testing-curve-fitting-and.html

International Congress: ISLAND EVOLUTION 150 YEARS AFTER DARWIN

150 Years after Darwin's On the Origin of Species, island evolution is entering a new phase. By habitat fragmentation, we humans create more and more islands, while at the same time, by transporting species from their native biomes, we remove the dispersal barriers that kept habitats isolated.

To explore the implications of this new era of island evolution, the National Museum of Natural History in Leiden, together with the Darwin Center for Biogeology in Utrecht, will organise an international congress on "Evolutionary islands 150 years after Darwin", to be held 12 & 13 February 2009 at the Museum Naturalis Leiden, the Netherlands.

The meeting will bring together traditional students of island biotas, experimental/ theoretical community ecologists, and evolutionary biologists, to explore the role of island-biological processes in a world in which the "island processes" of isolation and dispersal are being drastically altered.

Registration closes on January 28th, 2009. Abstracts for posters (A1 format, 59.4 x 84.0 cm) should be submitted to Jeremy Miller (miller@naturalis.nl) before December 15th, 2008.

For more information, scientific programme and registration:
http://www.naturalis.nl/darwin2009