J. Limnol., 2014; 73(s1): 161-170
DOI: 10.4081/jlimnol.2014.808

What can modern statistical tools do for limnology?

Andrew P. BECKERMAN*

Department of Animal and Plant Sciences, University of Sheffield, Western Bank, Sheffield S10 2TN, UK, and Instituto de Ciencias Ambientales y Evolutivas, Facultad de Ciencias, Universidad Austral de Chile, Campus Isla Teja s/n, Valdivia, Chile
*Corresponding author: a.beckerman@sheffield.ac.uk

ABSTRACT

Freshwater ecology and limnology is no stranger to experiments and data collected at large spatial and temporal scales. Nor is it a stranger to the molecular revolution. As our questions about distributions, abundances, structures and evolution grow in complexity with access to data, we need to stay aware of advances in tools for analysis and visualization. Here I review several advances in statistics and visualisation that might influence the future of limnology and freshwater ecology and evolution.

Key words: statistics, analyses, visualization.

Received: May 2013. Accepted: December 2013.

INTRODUCTION

Over the past decade, the power of personal computing has allowed numerous, previously *computationally intensive* statistical routines to become commonplace in ecology and evolution. his power on our desktops has allowed researchers to increase the spatial and temporal scale of analyses, increase the depth and rigour of analyses and improve the capacity to better and more formally link theory and data. All of this is complemented by major advances in graphics - both the ease of producing them and their capacity to present multiple forms of information have improved. This is good news. With global environmental change reaching across oceans and continents into ponds and streams, and with the current economic and political climate in many countries demanding research with *impact* the ability to visualise, model and analyse data at relevant spatial and temporal scales in a way that is accessible, reliable, repeatable and shareable, is critical.

Freshwater ecosystems are possibly one of the most under-appreciated ecosystems on the planet. While marine, agriculture and forested landscapes are tightly linked to a *grown and harversted* resource, freshwater systems provide a somewhat more passive resource - freshwater. Yet lakes, ponds, rivers and streams are central feature of landscapes around the world, and they are central features of economies and ecosystem services. They have offered science, government and economics some of the most amazing examples of how ecosystems operate. They have been central to revealing how agricultural and cultural toxins/runoff alter the functioning of whole communities (Carpenter *et al.,* 2008). The sensitivity of organisms in them have been central to revealing toxicity of numerous compounds in ecological contexts (Barata *et al.,* 2002; Preston, 2002; Shaw *et al.,* 2006; Wilding and Maltby, 2006; Asselman *et al.,* 2012). They have long been vital in assessing the stability and diversity of communities to large and small scale perturbations (Havel and Shurin, 2004 and see 2004 Special Feature in Limnology and Oceanography). And they have offered in sticklebacks and daphnia, two of the richest links between ecology and evolutionary/molecular biology (Hohenlohe *et al.,* 2010; Colbourne *et al.,* 2011, Mërila, 2014).

So, what can modern statistical tools do for limnology? On one hand, one might argue that we've done a pretty good job answering questions and making impact with what we have. It is a testament to the creativity of researchers in Limnology, and to the malleability of the systems, that this might be the case. One might also argue that if we need to resort to *new fangled* methods to answer our questions, the questions are not precise enough. If you can test your hypothesis with a *t*-test or some other *simple* test, some argue, you have missed asking the right question. But we know this is not true. To really understand whether modern statistical methods have anything to offer Limnology (and make this paper a bit longer), we need to be aware of what kinds of questions we want to be asking in Limnology. What challenges lie ahead for researchers in limnology that necessitate new, borrowed or revised statistical methods?

WHAT DO THE LAST 10 YEARS TELL US IS IMPORTANT?

In order to try answer this question, I used a rather popular, qualitative visualisation - wordclouds - of the last 10 years or so of freshwater research. While nice to look at, there are data underneath - the size of the word represents its frequency in the collection of text used to generate the cloud. I downloaded from ISI all articles from Limnology (2007- present) and from Freshwater Biology (2003 to present) and generated wordclouds from the abstracts and titles from each journal (Fig. 1). Aside from pointing out that work from Limnology reflects more on

sediment and *Daphnia* spp while work from Freshwater Biology reflects more on fish, the collection of words is broad. It includes, as we would expect, community ecology, ecotoxicology, issues about scale, space and time, *etc*. A focus on communities and assemblages stands out.

Such a retrospective analysis is never likely to foreshadow recently popularised concepts such as meta-community ecology or molecular ecology. But it also seems to reveal in the titles and abstracts little overt attention being paid to climate change, save for the words *temperature* and *responses* showing up rather infrequently. I think it is revealing that very little conceptual/theoretical detail can be gathered from these clouds - the algorithms do not detect the word *theory* very often; nor do they detect in this instance, phrases that might reflect conceptual advances rising in our field of research. So, from our titles and abstracts in these two journals, there is little sense of where limnology and freshwater ecology has really been: they do not reflect advances in ecological theory, developments in ecotoxicology, or the onset of molecular tools or remote sensing. This does not mean that these ideas have not been influential - just that our titles and abstracts are not reflecting this (I think that IS interesting).

ACCESSING DATA

While this *analysis* may not have been super revealing, it does highlight a new set of analysis tools available to researchers - tools to access multiple forms of data hiding in various places. Here, I used Web of Science data, exported into an XML format, imported into the R programming and statistical environment (R Core Team, 2013) using an XML

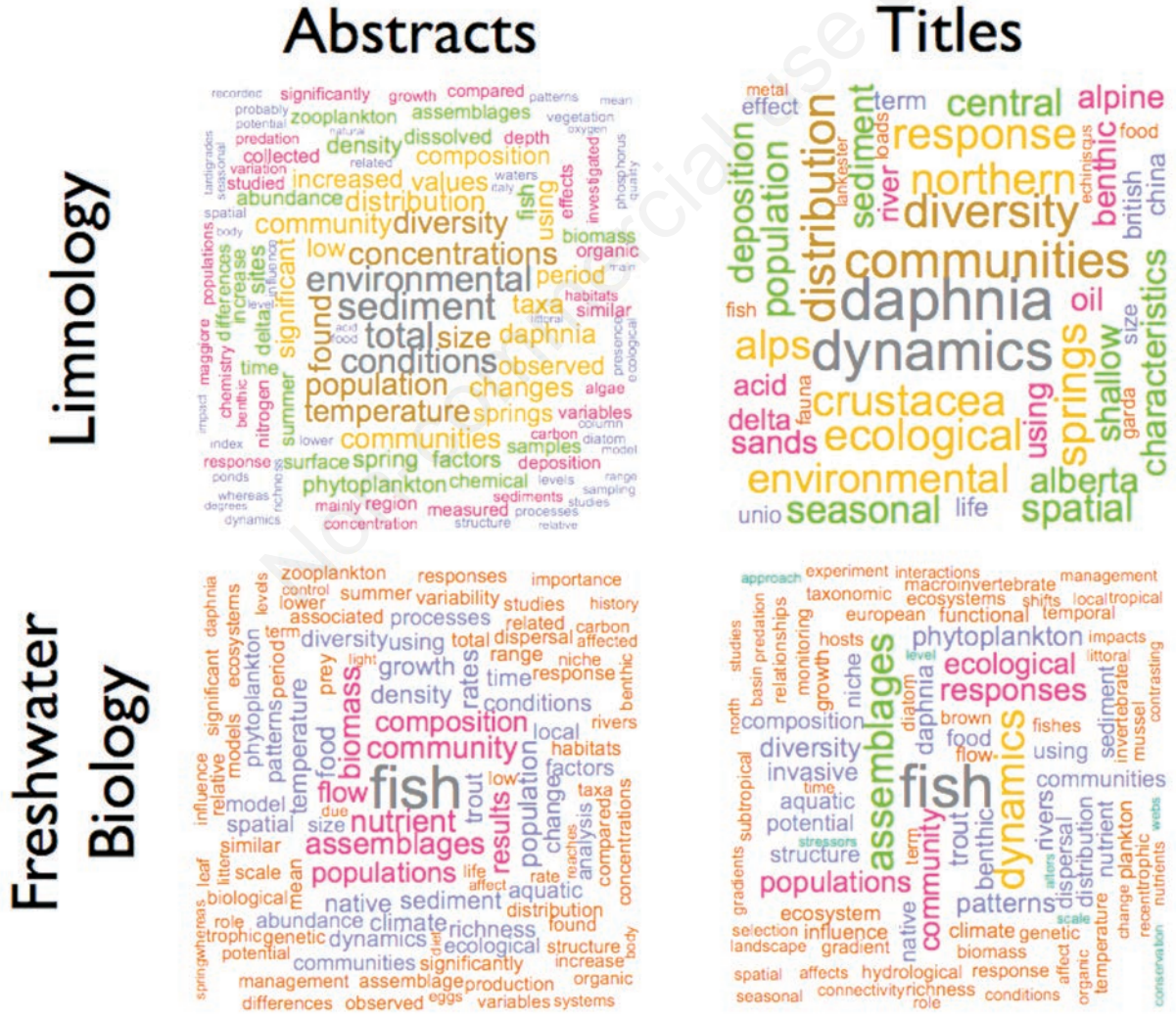

Fig 1. Wordclouds, visually representing the frequency of keywords found in the titles and abstracts of articles in the journal Limnology and in Freshwater Biology. The data are collected from ISI Web-of-Science, including articles from Limnology (2007- present) and from Freshwater Biology (2003 to present). The goal was to visualise themes of research permeating the field of freshwater ecology.

parsing package (Lang, 2013), further filtered by a the *tm* text mining package (Feinerer and Hornik, 2013), and then created the wordclouds with the wordcloud package (Fellows, 2013). This is also possible with GoogleScholar (https://code.google.com/p/google-scholar-word-cloud-r/). More generally, access to databases, the building of databases, the sharing of data and the analysis of cross-cutting, distributed information has long been seen as valuable - just look at the rise of meta-analyses in the last two decades. But increasingly mandated data-archiving requirements, associated databases such as Dryad (http://datadryad.org), and increasingly refined, freely available tools to access data (see http://ropensci.org), there is immense opportunity to further attempt to re-ask longstanding, theory driven questions we have about the effects of abiotic and biotic factors on the structure, complexity and diversity of lake ecosystems. I have no doubt that this will happen as various document formats become standardised, as organisations like ThomsonReuters, Google, PLoS and other publishers release and refine API's (Application Programming Interface) that allow access to various databases, as programming languages such as R, Python, Perl and Java become more popular, AND as PhD students emerge with more programming and technical skill.

Groups like ROpenSci (http://ropensci.org) are ensuring that we will see increasing effort and improved analyses focused on key questions that use data with historical/temporal or spatial footprints not normally associated with our research. Development of tools for several open source statistical and programming languages increasingly reflect the desire to access this type of information. For example, the new R-task view on WebTechnologies (http://cran.r-project.org/web/views/WebTechnologies.html) synthesizes a growing set of tools for parsing online data, managing access to data on the web and a rapidly increasing set of subject specific databases. And the Perl and Python programming languages have long had well established tools for *scaping* data from the web. Part of the justification for meta-analysis, re-analyses or expanded analyses, and their resulting synthetic interpretation, has always been to seek generality and thus increase the accessibility of our science to policy, by allowing us to emphasize repeated and important pattern. Tools to improve our access to new and varied sources data that allows such synthesis can only make our science more relevant to the public, policy makers, and dare I say, funders.

At this point, I refer readers to Box 1. Box 1 contains a set of links to resources that will help readers develop a more in-depth understanding of what tools and techniques are available for use. It includes websites, books and links to freely available resources, and there are sections associated with each of the discussion points that follow.

REPLICATION AND SCALE

The wordcloud exercise served to introduce numerous emerging tools for accessing data on the web that will increase opportunities for asking questions at new scales. But it is not only data from multiple studies that is motivating our assessment of pattern at larger spatial and temporal scales. We are designing experiments and surveys more and more at these large scales. This requires an awareness of statistical tools that accommodates new definitions of the replicate in our studies. Issues of replication and pseudo-replication have a long history of discussion in ecology (Hurlbert, 1984). We have always had questions about the generality of patterns among lakes and ponds, or between times. But when questions move to large spatial scales, for example, our unit of replication is often the location - and thus we need lots of them to answer questions at this scale - and sampling within the locations is considered the pseudo-replicate. There are similar issues with temporal data, quantitative genetic and phylogenetically structured data where repeatedly measured or related individuals are not independent.

These issues are collectively discussed in the context of non-independence. For questions at large spatial scales, we need many locations, with samples from within them. For questions of a temporal nature, we sample repeatedly the same individuals (or the same locations within sites over time). Spatial data, repeated measures of the same individuals or groups or sites, phylogenetic data and genetic signals are all sources of variation that we wish to understand and interpret. For example, repeated sampling of individuals is a central feature of estimating growth rates. We often sample numerous siblings within families or within genotypes (quantitative genetics) in order to estimate formally the heritability of traits central to understanding the process of natural selection. Phylogenetic information is required for questions cutting across different scales of biodiversity. And spatial information drives our capacity to generalise across biotic and abiotic changes in habitat. Whole lake, meso- and microcosm research (lakes, rivers, nets, bags, buckets and jars) remain mainstays of aquatic research - both for pure ecology and evolutionary ecology. In fact, many researchers turn to aquatic systems because of the capacity for these communities to allow replicated, randomised, blocked designs capturing several scales of variability.

The tools to deal with non-independence in data fall into the generic category of statistical modeling tools called random-effects models, hierarchical models or mixed-effects models (McCullagh and Nelder, 1989; Gelman and Hill, 2007). The following sub-sections centre on how to effectively manage spatial and temporal scale, and also quantitative genetic data with these tools. Understanding these tools and learning how to use them is critical in today's ecology. They are also a dynamic field of research in statistics, with some major debates about hypothesis testing. All the major statistical packages fit these models, but caution should be exercised (see below).

Linear Mixed Effects: the workhorse

General linear models are the class of models encompassing our standard friends ANOVA, ANCOVA and Regression. With these *basic* models, we make assumptions about the process generating the data (normal, Gaussian). Our response variable is some function of independent explanatory variables and their interactions, and after fitting such a model, we are left with some unexplained variation, and we assume this variation is normally distributed. In a mixed effects (random effects or hierarchical model) framework, we have a tool that allows us to relax the assumption about independence and partition formally the sources of variation over and above that associated with our treatments (fixed effects): in spatial, temporal and genetic data, the explanations of variation include variables

Box 1. Online and book resources.

Data access
ROpenSci: http://ropensci.com
RTaskView: http://cran.r-project.org/web/views/WebTechnologies.html

Mixed effects
Bates D, 2006. [R] lmer, p-values and all that. https://stat.ethz.ch/pipermail/r-help/2006-May/094765.html (R-help archived item by Prof. Douglas Bates, co-author of nlme and lme4 in R, on why p-values are hard in mixed models).
Bolker BM, Brooks ME, Clark CJ, Geange SW, Poulsen JR, Stevens MHH, White JS, 2009. Generalized linear mixed models: a practical guide for ecology and evolution. Trends Ecol. Evol. 24:127-135.
Crawley MJ, 2002. Statistical computing: an introduction to data analysis using S-PLUS. J. Wiley & Sons: 772 pp.
Pinhero JC, Bates DM, 2000. Mixed effects models in S and S-PLUS. Springer: 530 pp.
Zuur A, Leno EN, Walker N, Saveliev AA, Smith GM, 2009. Mixed effects models and extensions in ecology with R. Springer: 574 pp.

Generalised linear models (including survival models)
Faraway J, 2011. Functions and datasets for books by Julian Faraway. http://cran.r-project.org/web/packages/faraway/index.html
Crawley MJ, 2002. Statistical computing: an introduction to data analysis using S-PLUS. J. Wiley & Sons: 772 pp.
Hadfield JD, 2010. MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R Package. J. Stat. Softw. 33:1-22.
Harrell FE, 2001. Regression modeling strategies with applications to linear models, logistic regression and survival analysis. Springer: 571 pp.
McCullagh P, Nelder JA, 1989. Generalized linear models, 2. Chapman & Hall: 532 pp.

Bayesian and Bayesian MCMC methods
Gelman A, Carlin JB, Stern HS, Rubin DB, 2003. Bayesian data analysis, 2. Chapman & Hall: 696 pp.
Gelman A, Hill J, 2007. Data analysis using Regression and Multilevel/Hierarchical Models. Cambridge University Press: 625 pp.
Hadfield JD, 2010. MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R Package. J. Stat. Softw. 33:1-22.
McCarthy MA, 2007. Bayesian methods for ecology. Cambridge University Press: 296 pp.

WinBugs (http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml)
JAGS (http://mcmc-jags.sourceforge.net),
MCMCglmm: http://cran.r-project.org/web/packages/MCMCglmm/index.html

Graphics
Sarkar D, 2008. Lattice - multivariate data visualization with R. Springer: 268 pp.
Wickham H, 2009. ggplot2: elegant graphics for data analysis. Springer: 212 pp.

Molecular resources
Bioconductor (http://www.bioconductor.org).

Additional Wiki's
R-Phylogenetics Wiki - http://www.r-phylo.org/wiki/Main_Page
GLMM for ecologists and evolutionary biologists - http://glmm.wikidot.com

Other software and programming languages
Perl (http://www.perl.org)
Python (http://www.python.org)

Books and Notes worth reading centred on R (Note the Springer Use R! series).
Beckerman AP, Petchey OL, 2012. Getting started with R: an introduction for biologists. Oxford University Press: 160 pp.
Dalgaard P, 2004. Introductory statistics with R. Springer: 267 pp
Faraway JJ, 2002. Practical regression and Anova using R. Chapman & Hall: 213 pp.
Stevens MH, 2009. A primer of ecology with R. Springer: 388 pp.
Venables WN, Ripley BD, 2003. Modern and applied statistics with S. Springer: 497 pp.
Sarkar D, 2008. Lattice - multivariate data visualization with R. Springer: 268 pp.
Wickham H, 2009. ggplot2: elegant graphics for data analysis. Springer: 212 pp.
Hadfield J. Course notes for MCMCglmm. In MCMCglmm: http://cran.r-project.org/web/packages/MCMCglmm/index.html

(factors or covariates) that ARE independent (like treatments), but also variables, like spatial location or repeated sampling or families, that introduce non-independence.

Time, space and phylogeny

Repeatedly measured individuals or subsampled sites require that we formally recognize that each replicate introduces a source of non-independent variation. Traditionally, perhaps, we considered such effects by the concept of a random *block*. More generally in the terms of regression, we can consider that each replicate has its own intercept. This simple idea, when combined with a balanced and orthogonal design, has a very simple solution statistically where we can partition and estimate this variance and statistically formulate an F-test as a function of the variances and the degrees of freedom associated with the blocks, the groups defining our treatments and the random, residual variation.

However, our questions and experimental the designs are increasingly more complicated. Our questions about growth rates for example require that we not only represent each individual as a source of variation, but that we formally estimate the slope - the rate of growth - for each individual. Our questions about space often involve multiple, nested scales of sampling; our multiple watersheds contain multiple lakes in each of them, from which we sample multiple times.

Furthermore, we do not always succeed at a balanced design or perfectly orthogonal experiment. The tools to manage these more expansive questions and problems with design have long had a solution in a tool called Restricted (Residual) Maximum Likelihood. A computational extension of Maximum Likelihood methods, which themselves are, for simplicity, a computational generalization (see below in discussion of non-normal data) of least squares, REML allows us to estimate multiple sources of nested and non-nested variation in our data, and even when our experiments are unbalanced and not completely orthogonal (see Ovaskainen and Soininen, 2010).

Spatial variation (spatial autocorrelation) receives additional special attention in several statistics packages via extensions to linear mixed models. These extensions can specifically help evaluate the pattern of spatial autocorrelation in the residuals of your models, and actually fit variance (random) terms to capture these patterns. The excellent book *Mixed effects models in S and S-Plus* (Pinhero and Bates, 2000) is well worth a read on this topic along with more example driven works such as *Mixed effects models and extensions in ecology with R* by Zurr *et al.* (2009) and *Linear Mixed Models: a practical guide using statistical software* by West *et al*. (2007). Phylogenetic relatedness is captured similarly, but often leverages a class of model called Generalised Least Squares. The most popular, historical package - CAIC (Purvis and Rambaut, 1995) - has been ported and extended in R via the *caper* package (http://cran.r-project.org/web/packages/caper/vignettes/caper.pdf) and is complemented by many functions in the 'ape' package (http://cran.r-project.org/web/packages/ape/) all of which are tied together in a lovely wiki (http://www.r-phylo.org/wiki/Main_Page).

Finally, it is worth pointing out that questions about space and time were defined above in terms of random effects. However, the explicit analysis of time series and spatial data have very well developed tools; time series analyses have a rich history in econometrics and are equally available to ecology (Bjornstad and Grenfell, 2001) and increasingly valuable as longer and longer time series become available in our lake and pond communities (George and Harris, 1985; Thackeray *et al.,* 2013). The CRAN - Spatial task view (http://cran.r-project.org/web/views/Spatial.html) offers insight into the types of tools that are available and being developed, including interfaces with databases (including Landsat and Google Maps) and mapping programmes (ESRI/GRASS).

Again, I draw reader attention to Box 1 and Box 2. As noted above, Box 1 summarises a set of resources available to readers to gain more insight into methods. Box 2 highlights the value of the Task Views section on the Comprehensive R Archive Network. Whether you use R or not, this resource is invaluable as it contains explanations of several methods and via R, reproducible examples that work on most computing platform.

ARE WE NORMAL? GENERALISED LINEAR MODELS

The tools discussed above centre on non-independence. However, the second major set of assumptions in a typical general linear model (ANOVA, ANCOVA, Regression) centre on normality of the residuals. In ecology and evolution, there are several types of data where we *a priori* do not expect the residuals to be normal. Binomial or logisitic data, such as presence absence data in conservation, percentages of any sort, and sex-ratio data can now be treated explicitly via generalized linear models; there is no need for transformations. Additionally, count data, a core data type for conservation biology (*e.g*., species richness), is typically modelled using the poisson distribution. The major advances in dealing with binomial, poisson or any other data derived from other non-normal processes centre on the ease with which we can now fit these generalized linear models. These models allow estimation of effects associated with treatments where poisson, binomial and other *non-normal* data can be analysed with precision and ease - and without the frustration and risks associated with transformations (O'Hara and Kotze, 2010).

These data types present several problems - they are typically bounded in one way or another (*i.e*., counts are never less than 0; proportions are between 0 and 1) and the data have interesting mean-variance relationships. These models deal with these issues elegantly, offering precise and accurate inference and insight, over and above transformation methods (O'Hara and Kotze, 2010). For example, there are several new tools for dealing with

count data when there are more 0's than expected - the so called zero-inflated poisson. And parallel developments allow models to be fit to species richness data that simultaneously allow estimating the presence-absence process (binomial) and the species richness process (poisson). These are known as *hurdle* models. The pscl package in R (Jackman, 2011) has an excellent vignette and introduction to these tools. All statistical packages worth their programmers offer the possibility to fit generalised linear models, select the underlying distribution and make inferences. Again, Box 1 presents several resources for learning about and working with non-normal data.

Survival models

I think it is worth highlighting a class of model that is not used extensively throughout ecology, and perhaps even less so in aquatic ecology; that said, their importance in ecotoxicology cannot be understated, so there is a portion of our community very familiar with them. The models are commonly known as survival models, but are essentially models of the timing of events. They are variations on the generalised linear model, because the timing of events typically does not generate a normally distributed set of residuals/errors. There are parametric and non-parametric versions of these, and their importance to medical/epidemiological work means that they are constantly advancing. The key features of these models is that the response variable is a combination of information on the time an event occurs and the nature of that event. The latter could be, did it actually happen, or not, during my observation period. The value in this statement is that the distribution of event times is made up of known events, and information leading up to, for example, the last time we see an individual but before an event occurs. This abstract description of events can be infuriating, but it is key to think broadly. Event times can include the obvious - death and birth - but also might include hydrogeomorphic events (time to drying), breeding events, or foraging events.

The point here is that questions linked to when something occurs are the remit of these models. Frank Harrell's (2001) book *Regression modeling strategies with applications to linear models, logistic regression and survival analysis* is a great resource.

The generalised linear mixed model

The previous two sections introduced methods for dealing with either the non-independence in ones data, or with data arising from processes that generate non-normal error/residuals. Of course, there are methods to fit models that accommodate both situations: the GLMM - generalised linear mixed models. They are however very challenging. And while parametric methods are abundant, remember that there are debate about hypothesis testing linked to linear mixed models with normally distributed errors. One can imagine then that GLMM's are definitely worth being careful of. However, SAS, R, and Genstat/ASREML and Stata are the go-to programmes for fitting these models. Bolker *et al*. (2009) TREE review of GLMM's is a great starting point philosophically, but is now out of date with respect to package capabilities. The capacity to fit these models and make inference is improving with access to Bayesian methods for estimating GLMM's (see below) as highlighted by the work of Ovaskainen and Soininen (2010) in diatom presence - absence data in seven watersheds.

Multivariate data

Questions about multiple species, multiple traits or functional diversity remain significant features of research in lakes and ponds. Our research in these areas has long benefited from continually advancing methods of ordination (*e.g*., PCA, factor analysis, NMDS). Multivariate methods have been refined dramatically over the past decade, allowing the sophistication and complexity of our questions about biodiversity and for example, multiple traits in a life history, to mature.

Box 2. Task views from the Comprehensive R Archive Network.

Whether you use (or believe in using) R or not, the Comprehensive R Archive Network (CRAN) hosts an amazing resource tied to the >4500 packages available for use in data analysis and programming: Task Views (http://cran.r-project.org/web/views/). Task Views are curated selections of statistical and programming packages focused on questions and methods central to specific topics. For example, the Time Series Task View (http://cran.r-project.org/web/views/TimeSeries.html) is divided into Basics, Times and Dates and Classes - tools for handling this type of data - and then Forecasting, Frequency analysis, Decomposition and Filtering and Seasonality … *etc*. If you have done your background reading, and know a bit about why you want to use time series analyses, this resource, with packages, linked help files in pdf format, and example vignettes full of reproducible examples, is an online repository for learning about and taking advantage of some of the most advanced methods available.

Relevant to our ecological and evolutionary questions in aquatic systems, there are Task Views on Bayesian methods, Clustering, Differential Equations, Environmetrics (multivariate), Experimental Design, Genetics, Graphics, MetaAnalysis, Multivariate, Optimization, Spatial, SpatioTemporal, Survival and Times Series, *Phylogenetics, Especially Comparative Methods* and a very recent addition, Web Technologies and Services, for accessing databases and data services.

http://cran.r-project.org/web/views/

There are two very fruitful avenues of research relevant to freshwater ecology. First, for Multivariate Linear Models (*i.e*., MANOVA, but no mixed/random effects), there are several new tools being developed, largely in the *vegetation modelling*/*diversity*/*conservation* world that might offer limnology and aquatic ecology a very useful framework for conservation centred research. These methods draw on the rich history and range of tools for multivariate data ordination (PCA, CCA, MDS, *etc*.) and the increasing desire to embed this in a more formal hypothesis testing framework. One example is MVAbund (Multivariate Abundance; Wang *et al.,* 2012), an R package designed specifically for hypothesis testing with multivariate count data (site x species matrices) in an experimental framework. You can think of this as extending decades of developments centred on ordination methods designed to assess the correlation between ordination axes and environmental variables, such as found in the vegan package (Oksanen *et al*., 2013), and specifically the ordisurf() and envfit() functions. These were developed to estimate linear relationships between some factor and the distance matrix defining the ordination. MVAbund brings a multiple regression, multivariate, hypothesis testing framework to these problems (Wang *et al.,* 2012). As with all R packages, this one comes with great examples, vignettes and an associated published paper (Wang *et al.*, 2012), not to mention an hilarious video (www.youtube.com/watch?v=KnPkH6d89l4).

The second tool, a different philosophy as well, centres on Bayesian MCMC methods. Bayesian methods offer an alternative hypothesis testing framework (Gelman *et al.,* 2003; Gelman and Hill, 2007), and significantly, one for experimental situations where estimating variances (genetics, temporal variation spatial variation, *etc*.) are important (Ovaskainen and Soininen, 2010). Reflecting on previous sections, Bayesian MCMC methods are well suited to (generalized) (multivariate) mixed effects models. These methods are benefitting from our substantial desktop computing power, providing an alternative to p-value based inference, and facilitating the fitting of an extremely wide array of model types. If you find yourself having designed an experiment requiring hypothesis testing in a multivariate mixed model framework, you will probably want to learn how to use WinBugs (http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml), JAGS (http://mcmc-jags.sourceforge.net), and the MCMCglmm package in R (Hadfield, 2010), which is highly orientated to biological questions. Mark Kéry's book on WinBUGS (Kéry, 2010), Michael McCarthy's (2007) *Bayesian Methods for Ecology* and Jarrod Hadfield's *Course notes* associated with the MCMCglmm package in R are valuable assets.

Molecular data

Molecular biology is increasingly influential in ecology. We increasingly rely on molecular biology tools for species identification purposes, and our capacity to ask and re-ask questions about local adaptation with high resolution genomic (and other *omic* data) is growing. We are increasingly able to design experiments that with these types of data truly develop mechanistic understandings of how adaptation occurs, not to mention contributing to research focused on key question in evolutionary biology such as the stability of the genetic covariance matrix, genome evolution and rates of evolution (Hohenlohe *et al*., 2010; Mërila, 2014). This forum us too short to detail sufficiently what is on offer, but tools for managing, visualizing and analysing bioinformatics data is well served by the open source extension of R known as Bioconductor (http://www.bioconductor.org). This world of genomic data analysis is really served now by a body of programming savvy individuals leveraging a combination of programming languages ranging from Perl, Java and Python for data manipulation and parsing, to R for visualisation and analysis. These types of data sources typically leverage workflows that tie together the set-up and management of SQL like databases, Python/Java/Perl programmes that parse and often analyse data leading to more analysis and visualisation in R. Analysis and inference with molecular data is often associated with some of the same issues we introduced above; both pseudo-replication and response variables that are bounded in one way or another are features of molecular data. Bayesian methods, permutation methods, hierarchical models (mixed effects models) and various forms of *correction* for multiple testing (*e.g,* Bonferroni and False Discovery Rates) are central features of these developing methods.

Putting methods in context

Climate change, fisheries, species invasions, biodiversity and functional diversity drives a huge portion of research in aquatic communities. These questions require data at large scales, are multivariate and require these tools linked to space and time. The issues and tools discussed above - data management and statistical - are important in these applied realms. Research programmes in climate, fisheries and biodiversity are comprised of several types of data gathering, interpretation and analysis. They are comprised of i) a naturalist side - collecting, collating and sythesizing data on pattern; ii) a descriptive/hypothesis generating side - often using macroecological or multivariate/ordination tools to reduce the dimension of a high dimension dataset to further synthesize and visualise patterns (*e.g.,* spatially resolved climate or biodiversity data); and ii) a modelling side, where mathematical/statistical models are developed to both explain and predict the future. All three are vital, and together make clear that: a) the management and storage of data for access by others is vital; b) that statistical methods for describing and visualising high dimension data (multivariate) at large scales are incredibly valuable and typically drive hypothesis generation; c) that generating mathematical models provides testable theory (differential equations, partial differential equations); and d) fitting these models (*e.g*., state space modelling) to the

data at appropriate temporal and spatial scales completes a rather large circle.

Critically, as we enter an even more computer and data rich world, where the question by the next generation of scientists will be not whether, but how, to get access to data, the buzzword on the horizon is *reproducible research*. This is a catch-all term for a rather simple idea - data, analyses and results from publicly funded work should be reproducible. This requires that the data, code and results be accessible and available. I would assume that most readers understand what open source software means, and are at least familiar with Open Access publishing. It is certainly worth encouraging young and seasoned researchers in aquatic ecology to engage in the debate and where possible make tools, methods, data and results accessible and available. And the importance of this in the realm of climate change, access to freshwater and fisheries can not be understated.

On asking questions and making figures

The previous sections have focused on a few key advances that continue to mature the field of statistical analysis relevant to the scale of data we can collect in freshwater ecology. To finish, I want to highlight something that I and colleagues always teach: ask good questions, rely on theory to generate expectations in your data, replicate and randomise appropriately and then, once you have your data, make a picture before your analysis. With a good, formal figure and visualisation of your data, the analysis will follow. Always ensure that you understand the distinction between describing pattern (*i.e.*, generating a hypothesis) and analysing pattern (*e.g.*, testing a hypothesis). There is no harm in emphasizing just how critical it is to make a figure in advance of your analysis. If you have a clear idea what you were expecting - *i.e.*, how theory predicts the relationship between your data should look - make that figure BEFORE you embark on the

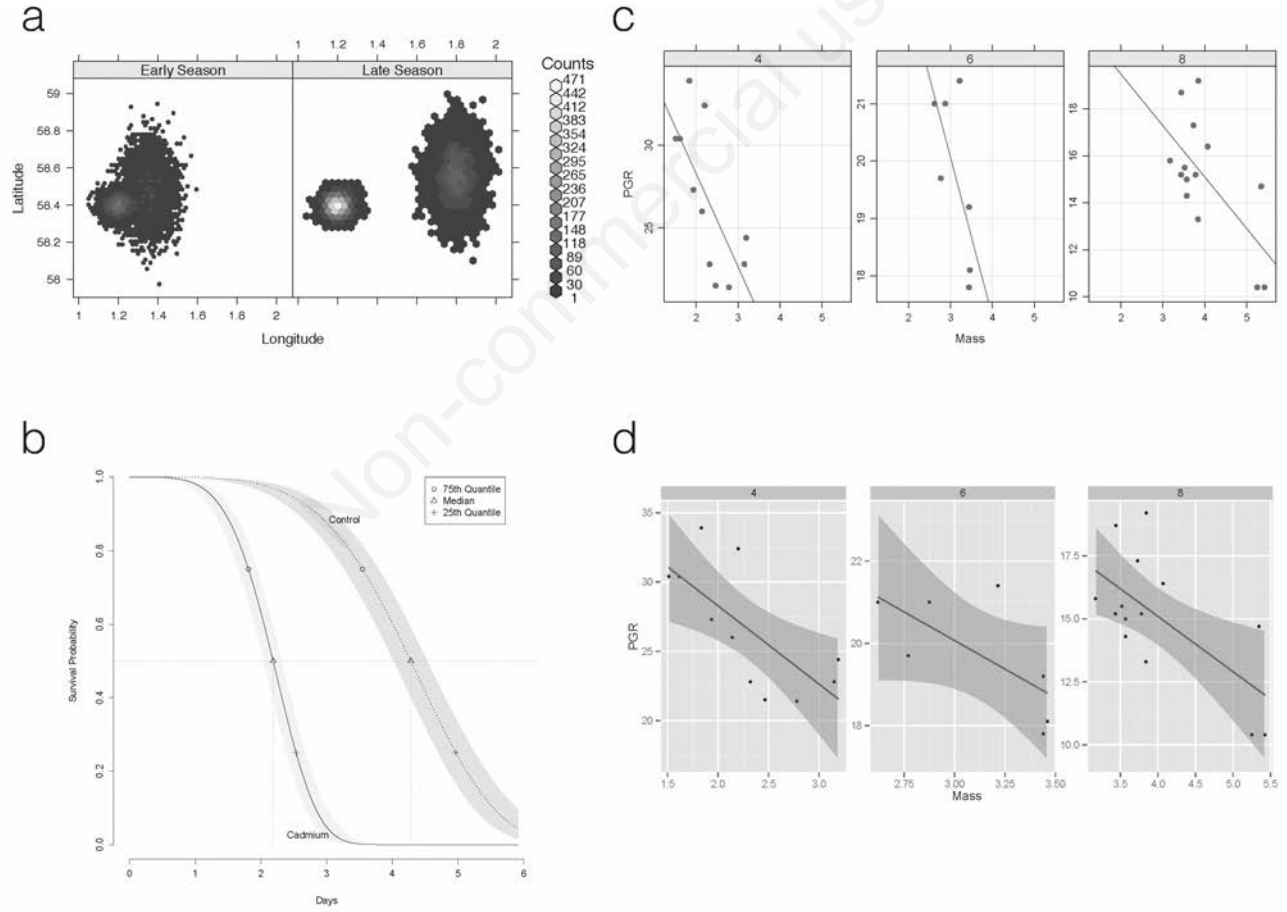

Fig 2. Examples of visualisation of data and models. A) a hexbin plot, revealing a 2-dimension spatial distribution of count data, from two different seasons, where the topology of abundance at sites is defined by the colour-scheme. B) The results of a hypothetical survival analysis comparing control and sub-lethal exposure to cadmium, demonstrating visualisation of confidence intervals and predicted percentiles on the survival curves. C and D) lattice (C) or facet (D) plots from the lattice and ggplot2 packages in R, revealing capacity for visualising multi-dimensional data, statistical fits and confidence bands in a compact and effective manner.

analysis. Visualization methods in a few modern statistical packages such as R, SPLUS, and Genstat have improved dramatically. One such improvement centres on lattice style graphics which allow structured, multivariate data to presented in very innovative and accessible forms, allowing researchers to see complex patterns in their data. here is an outstanding introduction to lattice graphics the lattice book (http://lmdvr.r-forge.r-project.org/figures/figures.html) by the lattice developer for R, Deepayan Sarkar (2008). The graphics language of ggplot2 (Wickham, 2009) is a rapidly emerging and very popular tool (http://ggplot2.org). Similar developments have followed in Genstat, SAS and STATA, as well as MATLAB, Python and other more formal mathematical and programming languages. It is worth remarking on how the Python community is developing as well as an integrated environment for data acquisition, manipulation, analysis and presentation (http://www.python.org and http://www.scipy.org).

My fundamental recommendation centres on making figures: the knowledge you gain of your data by making a figure that matches the theory/experiment/sampling programme you invested in pays major dividends. As the scale of our questions increases (either by looking at more, or looking in more detail), the need to visualise in advance and as a result of analyses can not be overlooked. When teaching bio-statistics, or working with PhD or masters students, we always require that they produce a figure before embarking on analyses (or consulting with us). Take some time to learn more visualisation tools and techniques. I would suggest using R, because the graphics are amazing quality and the process of having to ask (by writing down) for exactly what you want is one of the best exercises for getting to be familiar with your data. And if you've got a good idea of what you were expecting/hoping to see, this picture should set you on your way for a very straightforward analysis.

Fig. 2 a-d offer some insight into the open-source, visualisation tools available *via* R. In Fig. 2a I have created some spatially resolved presence - absence data on a phytoplankton bloom in an early and late season. Built in tools within the lattice (Sarkar, 2008) and hexbin package in R (Carr *et al.*, 2013) *bin* the counts (presences) at fine scales and present the data has hexagonal shapes that together generate a spatial representation of the density of the phytoplankton. This is a form of bivariate histogram useful for visualizing the structure in datasets with large n (Carr *et al.*, 2013). A similar resource is provided in the ggplot2 linbary (Wickham, 2009). Fig. 2b presents some simulated data on survival of a zooplankter in facing control and Cadmium treatments. Here, having fit a parametric survival model, a special form of the generalized linear model, built in tools within the *rms* package by Frank Harrell (2013) allow us to visualize clearly the fitted estimates of survival probability over time, with elegant, transparent 95% confidence intervals. I have also shown the median, 75th and 26th percentiles, easily calculated from the model and transferred to the figure. Fig. 2 c,d present lattice (Sarkar, 2008) and ggplot2 (Wickham, 2009) representations of the same data. In each case we are looking at simulated data on Population Growth Rate of several replicates of a zooplankter spread across a range of initial sizes (mass) of the starting propagules. Furthermore, the data were collected under three different conditions (Resource Levels=4,6,8). In each figure, we have the resource level specific data and linear regression of PGR *versus* starting mass, split by the treatments. There is a background grid provided in each, and in the ggplot2 implementation, confidence bands in transparent grey are also provided automatically around the panel specific linear regressions. The very simple message to drive home is that it is very possible to visualize complex data in a very effective manner that reflects the nature of a priori questions. Lattice style graphics, colour gradients, and transparent colours are all standard now in the advanced graphics modules in most statistical packages. Learning the syntax/language of the graphical tools pays major dividends for explaining your results. Programming is thus becoming increasingly important. For the novice, I do believe that R is an obvious choice for learning the basics of programming while re-enforcing the tenets of statistics and visualization (Beckerman and Petchey, 2012). Python, Perl and Java remain major forces in data management and in work with bioinformatics data. An important point, in light of the issues of public funding=public access, is to consider open source, reproducible research, and when possible open access. With R, Python and Perl, along with various open source database links, and associated outstanding statistical and programming communities offering loads of support, we have never been in a better situation to access tools for analysis and visualization of complex data.

CONCLUSIONS

Statistics and visualisation tools are always going to improve and or change. While it is easy to get caught up in statistical and analytic methods (there is a whole journal for that: Methods in Ecology and Evolution), the main advances in our science will still come from remembering to ask good questions, to rely on and develop theory to generate expectations, and to make good, informative figures that reflect you are *a priori* understanding. The analysis you need to perform will be crystal clear if you have a good visualization of your data.

REFERENCES

Asselman J, De Coninck DIM, Glaholt S, Colbourne JK, Janssen CR, Shaw JR, De Schamphelaere KA, 2012. Identification of pathways, gene networks, and paralogous gene families in daphnia pulex responding to exposure to the toxic cyanobacterium *Microcystis aeruginosa*. Environ. Sci. Technol. 46:8448-8457.

Barata C, Baird DJ, Mitchell SE, Soares A, 2002. Among- and within-population variability in tolerance to cadmium stress in natural populations of *Daphnia magna*: implications for

ecological risk assessment. Environ. Toxicol. Chem. 21: 1058-1064.
Beckerman AP, Petchey OL, 2012. Getting started with R: an introduction for biologists. Oxford University Press: 160 pp.
Bjornstad ON, Grenfell BT, 2001. Noisy clockwork: time series analysis of population fluctuations in animals. Science 293:638-643.
Bolker BM, Brooks ME, Clark CJ, Geange SW, Poulsen JR, Stevens MHH, White JS, 2009. Generalized linear mixed models: a practical guide for ecology and evolution. Trends Ecol. Evol. 24:127-135.
Carpenter SR, Brock WA, Cole JJ, Kitchell JF, Pace ML, 2008. Leading indicators of trophic cascades. Ecol. Lett. 11:128-138.
Carr D, Nicholas Lewin-Koh N, Martin Maechler M, 2013. hexbin: Hexagonal Binning Routines.
Colbourne JK, Pfrender ME, Gilbert D, Thomas WK, Tucker A, Oakley TH, Tokishita S, Aerts A, Arnold GJ, Basu MK, Bauer DJ, Cáceres CE, Carmel L, Casola C, Choi JH, Detter JC, Dong Q, Dusheyko S, Eads BD, Fröhlich T, Geiler-Samerotte KA, Gerlach D, Hatcher P, Jogdeo S, Krijgsveld J, Kriventseva EV, Kültz D, Laforsch C, Lindquist E, Lopez J, Manak JR, Muller J, Pangilinan J, Patwardhan RP, Pitluck S, Pritham EJ, Rechtsteiner A, Rho M, Rogozin IB, Sakarya O, Salamov A, Schaack S, Shapiro H, Shiga Y, Skalitzky C, Smith Z, Souvorov A, Sung W, Tang Z, Tsuchiya D, Tu H, Vos H, Wang M, Wolf YI, Yamagata H, Yamada T, Ye Y, Shaw JR, Andrews J, Crease TJ, Tang H, Lucas SM, Robertson HM, Bork P, Koonin EV, Zdobnov EM, Grigoriev IV, Lynch M, Boore JL, 2011. The ecoresponsive genome of *Daphnia pulex*. Science 331:555-561.
Feinerer I, Hornik K, 2013. tm: Text Mining Package. Available from: http://cran.r-project.org/web/packages/tm/index.html
Fellows I, 2013. wordcloud: Word Clouds. Available from: http://cran.r-project.org/web/packages/wordcloud/ index.html
Gelman A, Carlin JB, Stern HS, Rubin DB, 2003. Bayesian data analysis, 2. Chapman & Hall: 696 pp.
Gelman A, Hill J, 2007. Data analysis using Regression and Multilevel/Hierarchical Models. Cambridge University Press: 625 pp.
George DG, Harris GP, 1985. The effect of climate on long-term changes in the crustacean zooplankton biomass of lake Windermere, UK. Nature 316:536-539.
Hadfield JD, 2010. MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R Package. J. Stat. Softw. 33:1-22.
Harrell FE, 2001. Regression modeling strategies with applications to linear models, logistic regression and survival analysis. Springer: 571 pp.
Havel JE, Shurin JB, 2004. Mechanisms, effects, and scales of dispersal in freshwater zooplankton. Limnol. Oceanogr. 49:1229-1238.
Hohenlohe PA, Bassham S, Etter PD, Stiffler N, Johnson EA, Cresko WA, 2010. Population Genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. PLoS Genet 6:e1000862.
Hurlbert SH, 1984. Pseudoreplication And the design of ecological field experiments. Ecol. Monogr. 54:187-211.
Jackman S, 2011. pscl: classes and methods for R Developed in the Political Science Computational Laboratory. Stanford University.
Kéry M, 2010. Introduction to WinBUGS for ecologists: a Bayesian approach to regression, ANOVA, mixed models and related analyses. Associated Press: 320 pp.
Lang DT, 2013. XML: tools for parsing and generating XML within R and S-Plus. Available from: http://cran.r-project.org/web/packages/XML/index.html
McCarthy MA, 2007. Bayesian methods for ecology. Cambridge University Press: 296 pp.
McCullagh P, Nelder JA, 1989. Generalized linear models, 2. Chapman & Hall: 532 pp.
O'Hara RB, Kotze DJ, 2010. Do not log-transform count data. Meth. Ecol. Evol. 1:118-122.
Ovaskainen O, Soininen J, 2010. Making more out of sparse data: hierarchical modeling of species communities. Ecology 92:289-295.
Pinhero JC, Bates DM, 2000. Mixed effects models in S and S-PLUS. Springer: 530 pp.
Preston BL, 2002. Indirect effects in aquatic ecotoxicology: implications for ecological risk assessment. Environ. Manage. 29:311-323.
Purvis A, Rambaut A, 1995. Comparative analysis by independent contrasts (CAIC): an Apple Macintosh application for analysing comparative data.
R Core Team, 2013. R: a language and environment for statistical computing. R Foundation for Statistical Computing.
Sarkar D, 2008. Lattice - multivariate data visualization with R. Springer: 268 pp.
Shaw JR, Dempsey TD, Chen CY, Hamilton JW, Folt CL, 2006. Comparative toxicity of cadmium, zinc, and mixtures of cadmium and zinc to daphnids. Environ. Toxicol. Chem. 25:182-189.
Thackeray SJ, Henrys PA, Feuchtmayr H, Jones ID, Maberly SC, Winfield IJ, 2013. Food web de-synchronization in England's largest lake: an assessment based on multiple phenological metrics. Glob. Change Biol. 19**:**3568-3580.
Wang Y, Naumann U, Wright ST, Warton DI, 2012. mvabund - an R package for model-based analysis of multivariate abundance data. Meth. Ecol. Evol. 3:471-474.
West BT, Welch KB, Gałecki AT, 2007. Linear mixed models: a practical guide to using statistical software. Chapman & Hall/CRC: 353 pp.
Wickham H, 2009. ggplot2: elegant graphics for data analysis. Springer: 212 pp.
Wilding J, Maltby L, 2006. Relative toxicological importance of aqueous and dietary metal exposure to a freshwater crustacean: implications for risk assessment. Environ. Toxicol. Chem. 25:1795-1801.
Zuur A, Leno EN, Walker N, Saveliev AA, Smith GM, 2009. Mixed effects models and extensions in ecology with R. Springer: 574 pp.