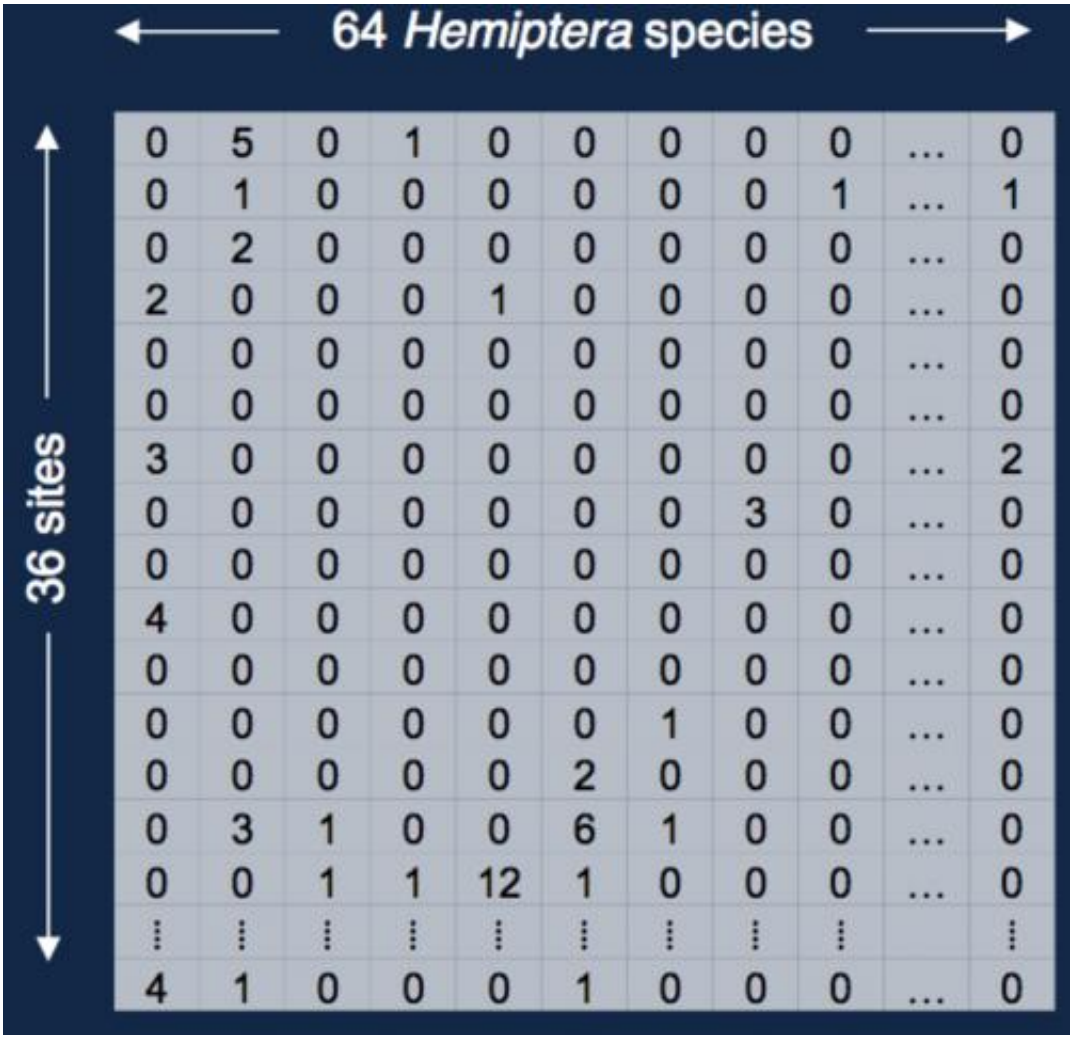
Older statistics methods being supplanted by more accurate ones

May 28 2014, by David Warton



It's an exciting time to be doing statistics. You heard me – statistics: exciting.

It often gets a bad rap, but stats is after all at the business end of the research process. When I've collaborated on studies of megafauna, leopard seals, police confessions, a new casino game or climate change effects on biodiversity, the point where researchers find out their results and have those "Eureka!" moments is more often than not in front of a computer rather than out in the field.

Now is an especially good time to be a statistician because the technological revolution over the past couple of decades has blown the field wide open – but despite this, some researchers continue to use outdated and inadequate statistical methods.

The sooner we can change this, the better.

When I started high school (around 25 years ago) computers looked like the one pictured right. They had 64kB memory. And this compressed digital image is more than 100kB, meaning that this poor computer doesn't have enough memory to look at this picture of itself!

Modern computers are thousands of times faster and can have a million times as much memory. This and related technological advances has changed data analysis in two main ways:

1. the increased capacity means that more computationally demanding methods of analysis are on the table that just weren't an option previously, so we can use more refined approaches to better answer important questions and even answer some questions not possible previously
2. technology has led to the collection of new data types, forcing statisticians to think in different ways to how they had before.

This has forced statisticians to think (amongst other things) about the challenging problem of high-dimensionality – when there are many more variables than subjects on which these variables were measured, leading to all sorts of methodological innovations.

The methodological advances over the past decade or two are hard enough to keep track of for someone with high-level training in statistics , so spare a thought for the applied scientist out there who has no clue what a sparse precision matrix is, let alone why you would want one or how an L1 penalty might help you get one.

Filling in the gaps

As an eco-statistician, an important part of what I do is to bridge the knowledge gap between statistics and ecology, which in one particular case has for some time been a huge gaping hole.

Ecologists often want to study whole communities rather than individual species (such as when looking for environmental impacts of human activities), and start by collecting data that look like the table shown left.

This particular dataset comes from a PhD student who has spent the past three years studying grassland communities of bugs (order Hemiptera) and how they change as temperature and rainfall changes (with a view to predicting how they might change as temperature and rainfall change in the future).

There are two important properties you should be able to see straight away:

1. lots of zeros (most species not found in most places)
2. lots of variables (many species recorded).

The first problem is relatively easy to deal with, but must be done with care. Having lots of variables ("high dimensionality") is a much more difficult problem technically.

In fact, if an ecologist walked into a statistician's office with this dataset in the 1980s there is a good chance they'd have been told to come back when they had fewer variables (hence a more refined and manageable problem).

Given the gap in the 1980s between what statisticians could offer and what ecologists needed, some ecologists quite rightly went off and made up their own methods, going in quite a different direction to the statistical literature of the time.

These involved some quite clever innovations to deal with their computational constraints (think 64kB), but also some compromises which I've recently found lead to some undesirable properties (partly because they don't deal with the lots-of-zeros property well).

In science, solving the problem isn't the end of the story. You then have to find ways to communicate your results and show people how your solutions can be useful to them.

This step is especially challenging in my case because the problematic approaches ecologists have been using since the 1980s are entrenched in the discipline, and are often even taught to ecology students in second year at university.

Even when I can show that in realistic settings their methods give wrong answers as much as 100% of the time, it is not a done deal for an ecologist to turn away from methods that have been standard in their discipline their whole career, and used countless times by them and their contemporaries. Never gonna give you up?

Roll in Rick Astley

This is where Rick Astley comes in. In order to convince ecologists to think differently about how they look for patterns in ecological communities, on YouTube (below) I compared the methods they use to Rick Astley's music – sounded like a great idea in the 80s, but these days, there are better options out there:

Rick was huge between mid-1987 and 1989, but his music dated very quickly (albeit with a resurgence following the Rick-Rolling phenomenon), and I've been making the case that the multivariate methods used in ecology have dated in much the same way, at least, from a methodological perspective.

On the whole this seemed to go over pretty well – the video is no Charlie Bit My Finger but 46 likes on the topic of multivariate analysis just might be a record! But the message here is that just as the technology available for doing science has changed, the technology for science communication has changed too.

I give several talks each year to other scientists, but in small groups, and over a 10-year period I will probably have presented to maybe 3,000 scientists. But I've managed to reach that many people in a single YouTube video that only took two days to write, record and edit.

Whereas a decade ago I felt a bit like a "lone voice" on modern approaches to modelling ecological communities, the good news is that this is changing, with exciting new ideas on the topic now emerging from a number of groups around the world. Happily, we are still making some inroads at the UNSW end too.

Source: The Conversation