Catlin-Groves, Christina L and Kirkhope, Claire L and Goodenough, Anne E and Stafford, Richard (2009). *Use of confidence radii to visualise significant differences in principal components analysis: Application to mammal assemblages at locations with different disturbance levels.* Ecological Informatics, 4 (3), 147-151. ISSN 15749541

Use of confidence radii to visualise significant differences in principal components analysis: application to mammal assemblages at locations with different disturbance levels

Christina L. Catlin-Groves, Claire L. Kirkhope, Anne E. Goodenough and Richard Stafford*

Department of Natural and Social Sciences, University of Gloucestershire, Swindon Road, Cheltenham. GL50 4AZ. United Kingdom

*Corresponding author

Tel: +44 (0)1242 714681

email: rstafford@glos.ac.uk

Abstract

Multivariate statistical analysis is a powerful method of examining complex datasets, such as species assemblages, that does not suffer from the oversimplification prevalent in many univariate analyses. However, identifying whether datapoints on a multivariate plot are clustered is subjective, as there is no determination of significant differences between the points and no indication of the level of certainty of those points. The validity of drawing such conclusions may therefore be considered suspect. This paper describes a method of bootstrapping calculated principal components to estimate a confidence radius, similar to confidence intervals in univariate techniques. Plotting 3D scatterplots of the principal components, with the size of the spherical point representative of the level of confidence of the estimate, gives a clear and visual indication of significant difference between the points – where spheres overlap there is no significant difference. We apply the technique to mammal assemblages at sites in Epping Forest (Essex, UK) that differ in the level of disturbance present and find that differences between some sites that appear large using traditional principal components analysis are actually not significantly different at the 95% confidence level, while other sites do differ significantly.

Key words

Principal Components Analysis; Bootstrapping; Confidence Intervals; Mammals; Community Assemblage; Disturbance

Introduction

Anthropogenic disturbance, for example through recreation activities, can influence the distribution and diversity of species, particularly for sensitive species such as some mammals (Cole and Landres, 1995; Gill et al., 1996). Use of univariate techniques such as measures of diversity and species richness to analyse changes in species assemblage is prone to species substitution (Rosenzweig, 1995; Gaston and Spicer, 1998). For example, if 'pest' or introduced species such as rats or mink displace native biota, including rare species such as water voles (Barreto et al., 1998), the results of univariate techniques may report the same level of diversity without reflecting important changes in the structure of the assemblage.

Multivariate techniques such as Principal Components Analysis (PCA) consider changes over an entire community assemblage, such that species substitutions show up as changes to the assemblage structure (McGarigal et al., 2000). PCA is a well-developed method for such analyses (e.g. Stemberger and Lazorchak, 1994), however, it is generally advised that the number of cases (e.g. sites at which samples were taken) is higher than the number of explanatory variables (e.g. the number of mammal species surveyed), normally at a ratio of 3:1 (Tabachnick and Fidell, 1989). Also, to ensure accuracy, the sample sizes should be relatively high: in most cases > 300 samples are needed for accurate results (Comrey and Lee, 1992). In much ecological research – particularly with respect to research on mammals where data are time consuming to collect – small sample sizes and high numbers of explanatory variables relative to data points are common (i.e. lots of species investigated at relatively few sites). While many statistical analysis programs allow calculations of PCA when these assumptions are violated (e.g. the 'prcomp' method in R), the results will not necessarily be reliable.

The purpose of this study is to develop a visual method of establishing whether clusters of data points are significantly different from one another, using the example of species assemblages at different study sites. The technique is analogous to the calculation of sample means and their respective confidence intervals in univariate analysis: if appropriate confidence limits of the different sample means overlap, no significant difference in mean values occurs (Schenker and Gentleman, 2001; Payton et al., 2003). In this case, confidence limits of the first three principal components are calculated by a bootstrapping exercise (similar to Yu et al., 1998). However, in our study, the assemblages at the study locations are plotted in three dimensions using the first three principal components as *x, y* and *z* coordinates, but the points are plotted as spheres, the size of the sphere based on the calculation of a confidence radius. This allows a clear visual interpretation of significant differences, rather than just investigating the precision of the estimate as demonstrated in previous studies (e.g. Yu et al., 1998). If spheres overlap, significant differences between the assemblages at the study locations are unlikely. If spheres do not overlap, then the assemblages at the respective sites can be considered significantly different. In this way, determining whether clusters constitute significant differences becomes less subjective than it is currently (Gabriel, 1971) and PCA becomes more analogous to an inferential statistical technique. Although this method would be useful in many situations, it is particularly relevant when the assumptions of PCA are not fully met and/or for small datasets.

Methods

Study site

The study was carried out at four different locations in Epping Forest, London during June 2008 (Table 1). These sites were similar in terms of habitat: mature semi-natural broadleaved woodland dominated by beech trees (verified using Phase 1 habitat surveying, which placed all 4 sites in the same habitat grouping: A.1.1.1).

Measuring Disturbance

Belt transects (160x100 m) were located at each site. Disturbances were recorded by surveying number of people to pass through the area and type of activities performed (jogging, cycling, dog walking, walking, picnicking and game playing – e.g. football). Relative values were given to the disturbance activities on a scale of 1-10 according to their perceived disturbance effects (see indications of disturbances of these activities in Cole and Landres, 1995; Gill et al., 1996; Delaney et al., 1999; Papouchis et al., 2001; Frid and Dill, 2002). Each site was given a disturbance value based on the number of individuals/groups multiplied by their corresponding activities.

Assessing mammal assemblages

A complete systematic sweep survey of each site (as defined by the belt transect) was performed to establish mammal assemblages on the basis of direct counts and indirect evidence (Sutherland, 1996). Timed point counts were used to survey larger mammals, while baited Longworth traps were used to assess small mammal communities (Barnett and Dutton, 1995). Indirect evidence was studied in the field where possible, although samples (e.g. scats and hair) were taken to aid further identification when required. Track-beds (3 per site) were set with builder's sand and placed along animal tracks and around burrows, dens and setts to determine active status and occupying species (Fletcher et al., 1990). Track-beds were moistened

(using a general fine mist garden sprayer) and reset at dawn, mid-afternoon (approximately 4pm) and dusk. The number of direct sightings and the amount of indirect evidence was used to provide an estimate of abundance. Caution was applied to these measures of abundance to ensure that individuals were not double-counted (Ross and Reeve, 2003).

Calculating principal components for the sites

Fourteen different mammal species were identified across the four sites, giving a cases:variable ratio of 2:7, rather than the normally required 3:1. Principal components were therefore calculated using the 'prcomp' function in R (R Core Development Team, 2007), using the method described by Crawley (2007), since the cases:variable ratio precluded the use of the 'princomp' function. Variability between sites was large and so data were centred and scaled (as per the guidelines of Varmuza and Filzmoser, 2009) using the options in the 'prcomp' function.

Bootstrapping the principal components

For each site, a frequency distribution was set up whereby each individual of each species was set as a distinct data entry (as an example, species 1 (fox) where three individuals were identified, would be represented by 1, 1, 1 – see examples of actual data in supplementary material). From these data, 15 samples were taken with replacement, to obtain a sample of the mammal assemblage present at the site (i.e. the probability of obtaining a given species in a sample does not change for each sample, regardless of the number of each species already found). Thus, the probability of a given mammal species occurring in the sample at a given site was proportional to the abundance of the species directly observed or observed through indirect evidence

(tracks, scats etc.) at the site. The first three principal components of each sampling exercise were stored for each site, and the process repeated 10,000 times.

For each replicate run of the bootstrapped principal components (where *n* always equalled 15; see above), the full dataset for all four sites was also analysed, essentially creating eight points or sites in each replicate run. By calculating a vector to transform each point from the full dataset back to its corresponding point when calculated without the additional bootstrap points (equation 1), and then applying the same vector to the bootstrap points (equation 2), the variability in the bootstrapped points is restricted to variation between differences in the placement of points on the initial principal component axes, and not variation between both the placement of points and alignment of principal component axes. Accordingly, only the variability inherent in the actual data is included; variability is not increased as a facet of the bootstrapping analysis.

$$v_{[x,y,z]} = I_{[x,y,z]} - i_{[x,y,z]} \quad [1]$$

$$B_{mod[x,y,z]} = B_{calc[x,y,z]} + v_{[x,y,z]} \quad [2]$$

where *v* is the vector, *I* is the initial full data point calculated without the addition of the bootstrap points, *i* is the full data point calculated along with the bootstrap points, B_{mod} is the bootstrapped point modified by the vector and B_{calc} is the bootstrap point calculated directly by PCA.

Essentially this process is the same as rotating the axis from each replicate run so that the principal component axes align with those of the original data (i.e. the original data in the absence of bootstrap data). Applying this vector also accounted for

the arbitrary sign applied to the magnitude of the principal component (during replicates on identical datasets, the value of a point on a principal component axis could be assigned as 1 or -1). The vector transformation eliminated this problem unless the sign (+ or -) of the full dataset differed from the sign of the bootstrapped dataset for the same point. If this was the case, the magnitude of the vector in this dimension was ~ 2 x that of the magnitude of the value of the full dataset point. To account for this problem, if the magnitude of the vector exceeded 1.2 x that of the magnitude of the value of the full dataset point, the magnitude of the vector in this dimension was calculated by adding the two points (equation 3) and then subtracting the calculated bootstrap value from the vector (equation 4).

$$v_{[x,y,z]} = I_{[x,y,z]} + i_{[x,y,z]} \quad [3]$$

$$B_{mod[x,y,z]} = v_{[x,y,z]} - B_{calc[x,y,z]} \quad [4]$$

The value of 1.2 x the magnitude as the demarcation between equations 1 and 3 being applied was essentially arbitrary, but needed to be << 2 to ensure appropriate detection of occasions where different signs had been applied arbitrarily to the principal components. Results were robust to changes in this value between 1.0 and 1.5 using the current dataset.

The mean of each replicate of the vector-transformed principal component was calculated and 95% confidence limits were calculated by excluding the highest and lowest 2.5 % of the values (Crawley, 2005). Upper and lower confidence intervals for all three of the stored principal components were averaged to give a confidence radius. The mean values of the principal components for each site were plotted in 3 dimensions and the confidence radius indicated the size of the sphere.

Plots were made using the RGL library and rgl.sphere function for R (Adler and Murdoch, 2008). Both the R code and the data files are included as electronic supplementary material.

Results

Assessing mammal assemblages and disturbance

The number of mammal species found in each site varied between 6 and 12 (Table 1), with a maximum of 57 individuals (site 4) and a minimum of 14 individuals (site 2) (Table 2). The estimated value of disturbance for each site varied by a magnitude of 27, clearly suggesting far more disturbance present at some sites using the index devised using *a priori* knowledge (Table 1). Accordingly, separating the sites on the basis of anthropogenic disturbance was justified.

Principal components for the sites

The traditional biplot for the four sites, plotted using the first two principal components (accounting for 90.1 % of the variance, the first three principal components accounting for 100 %) shows the apparent clustering of sites 2 and 3, while sites 1 and 4 appear separate from both sites 2 and 3 and from each other (Figure 1). The species in the mammal assemblages that most account for the differences between sites are also indicated (Figure 1). Differences between the principal components and the bootstrapped values (below) can be seen in Table 3.

Bootstrapping the principal components

The mean values of the first three principal components for each study site, along with the upper and lower confidence limits and confidence radius, are given in Table

3. Plotting each site by its first three principal components, and associated confidence radius, demonstrates overlap between many sites (Figure 2a). Comparisons between the most disturbed and least disturbed sites (sites 3 and 4) demonstrate no significant differences in mammal assemblage, in fact these are the most similar sites with site 3 and its confidence radius is entirely subsumed by the confidence radius of site 4. While no sites are significantly different from site 4, significant differences do occur between all other possible comparisons (Figure 2 b-d) at the 95% confidence level. These results differ from a conventional interpretation of the initial PCA (Figure 1), which suggested most sites were distinct from each other, and if groupings were likely to occur, they would be between sites 2 and 3.

Discussion

Disturbance may affect community structure in several ways. For example, urban adapters and exploiters such as foxes, hedgehogs, squirrels and rats are likely to be attracted to disturbed areas due to litter, which may thereby increase the local population (McKinney, 2002). There may also be a concurrent decease in sensitive mammals such as badgers, which are not particularly adapted to urbanised areas (Harris, 1984). In the areas of higher disturbance at Epping Forest there was a greater number of scavengers and urbanized species, such as rats (*Rattus* spp.) and hedgehogs (*Erinaceus europaeus*), which did not appear to be present at less disturbed sites, as well as squirrels (*Sciurus carolinensis*) that occurred at a higher abundance at the more disturbed sites compared to the less disturbed sites. In contrast to this, and true to expectation, there was no evidence of badgers (*Meles meles*) at the disturbed sites, while the undisturbed sites showed recognisable secondary signs of badgers.

Whilst the results of this study did not differentiate between assemblages at sites with the biggest differences in disturbance, there are a number of limitations that should be considered. Only a few small mammal species were included in the study due to lack of sufficient captures in the field; evidence of small mammals is also difficult to identify using techniques such as track beds. This limitation can also influence the bootstrapping technique used in this study. If species were not identified in the field, then they could not be found in the bootstrapped sample. Missing a number of small mammal species, that may have been present, from some sites, but not from others, would be likely to result in false significant differences being found between sites. That these type II errors did not occur in the study indicates that the technique is relatively robust to these potential problems of small sample size.

In many cases, in ecological research, the assumptions of statistical analysis techniques are not met (Underwood, 1996). In particular, multivariate techniques such as PCA may not be accurate if assumptions are violated, and there is little way of telling if distinct clusters are really significantly different from each other without separate secondary inferential statistical analysis (Shaw, 2003). This study demonstrates that in the case of small samples, especially with many explanatory variables, conventional PCA can be misleading and in fact significant differences between sites that may initially seem distinct may not differ significantly, and equally, those sites that appear clustered may in fact show significant differences. This study, therefore, provides the basis of a new statistical technique to visually investigate multivariate results in a manner that deals with the realities of ecological data. Essentially the technique provides an element of quantification to what is normally solely a descriptive process. These techniques could provide important information in the fields of population, community and conservation biology by providing a simple

method of detecting differences (or similarities) in species assemblages. This method could also be applied more widely, being of potential use wherever determining whether or not data points produced by PCA are significantly clustered is important, including where data are limited.

References

Adler, D. and Murdoch, D., 2008. rgl: 3D visualization device system (OpenGL). R package version 0.77. http://rgl.neoscientists.org

Barnett, A. and Dutton, J., 1995. Expedition Field Techniques: Small Mammals. Royal Geographical Society, London, U.K.

Barreto, G.R., Rushton, S.P., Strachan, R. and Macdonald, D.W., 1998. The role of habitat and mink predation in determining the status and distribution of declining populations of water voles in England, Animal Conservation 1, 129–137.

Cole, C. and Landres, P., 1995. Indirect effects of recreation on wildlife. In: R. Knight and K. Gutzwiller (Editors): Wildlife and recreationists – Coexistence through management and research. Island Press, Washington DC. pp. 183-202.

Comrey, A.L. and Lee, H.B., 1992. A First Course in Factor Analysis (2nd Edition). Lawrence Erlbalm Associates, Hillsdale, N.J.

Crawley, M.J., 2005. Statisitics: an Introduction using R. Wiley, Chichester, U.K.

Crawley, M.J., 2007. The R Book. Wiley, Chichester, U.K.

Delaney, D., Grubb, T., Beier, P., Pater, L., and Reiser, H., 1999. Effects of helicopter noise on Mexican Spotted Owls. Journal of Wildlife Management, 63, 60-76.

Fletcher, W., Creekmore, T., Smith, M. and Nettles, V., 1990. A field trial to determine the feasibility of delivering oral vaccine to wild swine. Journal of Wildlife Diseases. 26, 502-510.

Frid, A. and Dill, L., 2002. Human-caused disturbance stimuli as a form of predation risk. Conservation Ecology. 6, 1:11.

Gabriel, K.R., 1971. The biplot graphic display of matrices with application to principal component analysis. Biometrika. 58, 453-467.

Gaston, K.J. and Spicer, J.I., 1998. Biodiversity. Blackwell Science, Oxford, U.K.

Gill, J., Sutherland, W.J. and Watkinson, A., 1996. A method to quantify the effects of human disturbance on animal populations. Journal of Applied Ecology. 33, 786-792.

Harris, S., 1984. Ecology of urban badgers (*Meles meles*): Distribution in Britain and habitat selection, persecution, food and damage in the City of Bristol. Biological Conservation. 28, 349-375.

McGarigal, K., Cushman, S., Stafford, S.G., 2000. Multivariate Statisitics for Wildlife and Ecology Research. Springer-Verlag, New York, N.Y.

McKinney, M., 2002. Urbanisation, biodiversity, and conservation. Bioscience. 52, 883-889.

Papouchis, C., Singer, F., & Sloan, W., 2001. Responses of desert bighorn sheep to increased human recreation. Journal of Wildlife Management. 65, 573–582.

Payton, M.E., Greenstone, M.H. and Schenker, N., 2003. Overlapping confidence intervals or standard error intervals: What do they mean in terms of statistical significance? Journal of Insect Science, 3, 1-6.

R Development Core Team, 2007. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Ross, C. and Reeve, C., 2003. Survey and census methods: population distribution and density. In: J.M. Setchell and D.J. Curtis (Editors), Field and Laboratory Methods in Primatology: A Practical Guide. Cambridge University Press, Cambridge, U.K.

Rozenzweig, M.L., 1995. Species Diversity in Space and Time. Cambridge University Press. Cambridge, U.K.

Schenker, N. and Gentleman, J. F., 2001. On Judging the Significance of Differences by Examining the Overlap Between Confidence Intervals. The American Statistician, 55, 182-186.

Shaw, P.J.A., 2003. Multivariate Statistics for the Environmental Sciences. Arnold, London, U.K.

Stemberger, R.S. and Lazorchak, J.M., 1994. Zooplankton assemblage responses to disturbance gradients. Canadian Journal of Fisheries and Aquatic Sciences. 51, 2435-2447.

Taberchnick, B.G. and Fidell, L.S., 1989. Using Multivariate Statistics. Harper Collins, New York, N.Y.

Sutherland, W.J. 1996., Mammals. In: W.J. Sutherland (Editor) Ecological Census Techniques – A Handbook. Cambridge University Press. Cambridge, U.K. pp. 260-280.

Underwood, A.J., 1996. Experiments in Ecology: their Logical Design and Interpretation using Analysis of Variance. Cambridge University Press. Cambridge, U.K.

Varmuza, K. and Filzmoser, P., 2009. Introduction to Multivariate Statistical Analysis in Chemometrics. C.R.C Press, Danvers, M.A.

Yu, C.C., Quinn, J.T., Dufournaud, C.M., Harrington, J.J., Rogers, P.P. and Lohani, B.N., 1998. Effective dimensionality of environmental indicators: a principal component analysis with bootstrap confidence intervals. Journal of Environmental Management. 53, 101-119.

Table 1. The location of each study site, with its respective mammal species richness and calculated disturbance value.

Site	Latitude	Longitude	Number of species	Disturbance
1. North Long Hills	0:01:53E	51:38:52N	8	19
2. Rangers Road	0:01:25E	51:38:07N	6	8
3. Wellington Hill	0:02:11E	51:39:54N	6	4
4. Pillow Mounds	0:02:21E	51:39:59N	12	109

Table 2. Abundance of each species found at the four sites.

	Site 1	Site 2	Site 3	Site 4
Fox	4	7	6	7
Rabbit	6	2	1	22
Hare	1	0	0	2
Stoat	2	0	0	1
Weasel	0	0	0	2
Rat	0	0	0	1
Squirrel	6	1	7	9
Fallow deer	5	0	2	2
Chinese water deer	0	0	3	5
Muntjack deer	4	0	0	3
Roe deer	1	1	0	0
Hedgehog	0	0	0	2
Bank vole	0	0	0	1
Badger	0	3	2	0

Table 3. The first three principal components from the full dataset and the mean values and upper and lower confidence intervals of the bootstrapped principal components. Confidence radii for the bootstrapped values are also given.

	SITE 1			**SITE 2**			**SITE 3**			**SITE 4**		
	PC1	**PC2**	**PC3**	**PC1**	**PC2**	**PC3**	**PC1**	**PC2**	**PC3**	**PC1**	**PC2**	**PC3**
Full data	-0.36	3.12	0.11	-2.55	-1.35	1.67	-1.19	-0.93	-1.57	4.11	-0.84	0.28
Bootstrapped (mean)	0.77	2.00	0.20	-2.57	-1.50	1.17	-1.44	-0.50	-1.71	-2.07	0.32	-0.85
Upper CL	1.55	3.69	3.34	-1.92	-0.59	2.56	-0.95	0.63	0.18	-0.35	2.17	4.26
Lower CL	-1.02	0.07	-3.16	-3.21	-2.42	-0.54	-2.05	-1.46	-3.04	-3.26	-2.04	-3.48
Confidence radius			2.11			1.03			1.00			2.48

Figure 1. Biplot of the first two principal components created using mammal assemblages at the study sites. Sites 2 and 3 appear to form a cluster, where as 1 and 4 appear to form distinct points. The mammal species indicated by the arrows indicate the main differences between the sites (i.e. site 2 and 3 largely separated from the other sites by the presence of badgers, site 1 separated by the presence of fallow deer, and site 4 showing a number of small mammal species present). The length of the arrows indicate the eignvector loadings, in this case all approximately equal at xxx.

Figure 2. Three dimensional principal component plots with 95% confidence radii. Numbers on spheres indicate the sites they represent (a) all of the sites – indicating overlap between sites 1 and 4, sites 2 and 4 and sites 3 and 4 (note site 3 and its confidence radius is subsumed by the confidence radius of site 4). (b - d) significant differences occur between sites when there is no overlap of the respective confidence radii.