**University of Alberta**

# How many fossils, how to count them, and where to collect them: Examinations of the most appropriate methods for community paleoecological research

by

Frank Louis Forcino

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Earth and Atmospheric Sciences



Fall 2013
Edmonton, Alberta

*Abstract*

Paleoecology researchers have employed an array of methods for collecting, counting, and identifying fossil data; no standard protocol exists for conducting community paleoecological research. The lack of a standard protocol could lead to inaccurate conclusions. In addition, paleocommunity research is labor- and time-intensive, as it requires expertise in multiple taxonomic groups and geological sub-disciplines. Therefore, it is important not to over-sample (individuals per sample or number of total samples), because an increasingly large sample size and number of samples will eventually result in diminishing returns in terms of improving any pattern revealed by the data. Resources should be allocated appropriately to learn as much as possible about the Earth's history.

Here, I examined (1) the spatial and temporal resolution of fossil sample collection, (2) the counting methods most appropriate for community paleoecological research (abundance or biomass), (3) the groups of fossil organisms that should be examined in order to gain an accurate picture of past ecosystems, (4) the taxonomic level of identification, and (5) sample size (number of individual fossil specimens collected per sample). The ultimate goal of this research is to provide a set of "best" or most accurate methods for use in community paleoecological research.

I found that a sample size of 50 is sufficient for community paleoecological research that employs multivariate statistical techniques. This value is supported more definitively when using fossils (30 datasets), but is still supported using 44 modern datasets. In addition, I demonstrate that fewer lateral samples are required when conducting community paleoecological research at relatively greater temporal scales. These data could potentially allow researchers to save time and money. Research efforts and resources can be focused gaining a greater number of samples per study or conducting additional studies.

There are areas where greater paleontological resources should be allocated. (1) Genus (or species) identification is required for an accurate representation of paleocommunities. (2) Whenever possible, researchers should tally the abundance of fossil taxa in concert with a biomass proxy (i.e., point counts). (3) In addition, researchers should examine all available taxa, as opposed to single taxonomic groups, such as only brachiopods.

*Table of Contents*

*List of Tables*

*List of Figures*

## Chapter 1. Introduction

Ecological research is important for gaining a complete understanding of biological and environmental systems. However, modern ecological research is limited to human time scales (tens to thousands of years). The fossil record is rich with information about past environmental and biological events (Sepkoski, 1981; Raup and Sepkoski, 1982; Bambach, 1993; McGhee, 1996; Bambach et al., 2004; Clapham and Bottjer, 2007; Stafford and Leighton, 2011; Tyler and Leighton, 2011). Due to the current biodiversity crisis and extinction event, it is important to be able to fully understand the changes occurring to the world's ecosystems (Hughes et al., 1997; Ceballos and Ehrlich, 2002; Dirzo and Rave, 2003; Pereira, 2010). Studying events from throughout the Phanerozoic can provide unique insight into changes that are currently occurring or that may occur in the near future (Barnosky et al., 2011).

By assessing ecosystem variation using complete communities of organisms (groups of organisms that directly or indirectly interact for resources), researchers can obtain higher-resolution examinations of environmental and ecological changes through time and space because the characteristic and niche overlaps among taxa are often more finely resolved than each individual taxon's characteristic or niche. Community paleoecology utilizes complete fossil assemblages to determine the mechanisms of spatiotemporal ecological and environmental variation, aiding in the pursuit of the processes that structure ecosystems and the causes of ecosystem collapse and extinction (Jablonski, 1998; Olszewski and Patzkowsky, 2001; Bonelli et al., 2006; Clapham et al., 2006; Clapham and James, 2008; Heim, 2009). Paleocommunities (herein referring to marine invertebrate fossil assemblages used to infer environmental or ecological gradients) provide a wealth of information regarding taxonomic interactions and environmental tolerances on local scales, and elucidate how these processes scale up to regional, continental, and global scales (Bambach, 1993; Kowalewski et al., 2002; Bambach et al., 2004; Clapham and Bottjer, 2007).

Researchers of community paleoecology have employed an array of methods for collecting, counting, and identifying fossil data; there is no standard protocol for conducting community paleoecological research. It is important to ensure that the methods in any particular study are producing the most accurate results. Furthermore, if researchers wish to combine data to conduct larger-scale studies and meta-analyses, standardized collecting protocols are needed. In addition, paleocommunity research is labor- and time-intensive, as it requires expertise in multiple taxonomic groups and sub-disciplines of geology to develop a stratigraphic and biological framework (Kowalewski et al., 2002; Forcino et al., 2010a). Therefore, it is important not to over-collect, because an increasingly large sample size or number of samples will eventually result in diminishing returns in terms of improving any pattern revealed by the fossil data. A paleoecologist's time and resources may be better spent acquiring additional samples to supplement other types of statistical power (Bennington, 2003; Zambito et al., 2008).

The specific methodological protocols I examine here are (1) the spatial and temporal resolution of sample collection, (2) counting methods (abundance versus biomass), (3) the groups of organisms that should be examined, (4) the taxonomic level of identification, and (5) sample size (number of individual fossils specimens collected per sample). The ultimate goal of this research is to provide a set of "best" methods for use in paleocommunity research. Even if a "best" set of methods cannot be determined (or does not exist), it is important for researchers to know whether two sets of methods can produce different paleocommunity results. If two sets of methods applied to the same series of paleocommunities produce different paleocommunity results, researchers may want to explore all possible methods and base any conclusions on the methods that are most suitable to the specific type of question being explored.

Throughout all of the research in this thesis, to compare the results of the various methods, I use statistical techniques that are commonly used in ecology, specifically for inspecting and analyzing a series of samples that each contains some number of taxa. Below, I describe these statistical methods and why they are employed.

1. Statistics

To extract the most information from analyses of the changes in life through time, paleobiologists often employ advanced statistical techniques (Olszewski and Patzkowsky, 2001; Kowalewski et al., 2002; Bonelli et al., 2006; Clapham et al., 2006; Clapham and James, 2008; Heim, 2009). These statistical techniques give researchers insight into the ecological intricacies that have typified the history of life (Maurer, 1999; Leibold et al., 2004). Univariate statistical methods (e.g., correlations), visual assessments (e.g., comparing graphs of the relative proportions of taxa), and other ecological measures (e.g., evenness) of the taxonomic compositions of fossil assemblages are important means of comparing multiple samples (Staff and Powel, 1999; Bennington, 2003). However, these types of comparisons are difficult when examining a large number of fossil samples with varying abundances of many different taxa, as is typical of the fossil datasets. Multivariate statistics overcome these limitations by providing efficient means of assessing similarities, differences, and changes among a large number of fossil samples through time and across space (Mantel, 1967; Jackson, 1995; Legendre and Legendre, 1998; McCune and Grace, 2002; Legendre, 2005; Vinatier et al., 2011).

The term "community" can be used in a number of ways. In general, a community is a group of organisms that live together and interact ecologically. In the literature, "community" is often used as a shorthand referring to a smaller sampled portion of a full community: a "community" may be a sampled plot of grasses within a larger prairie sample area, or a collection of fossil marine invertebrates from one stratigraphic bed within a larger span of a few million years (Olszewski and Patzkowsky, 2001; Leibold et al., 2004; Ricklefs, 2008; Forcino et al., 2010). A paleocommunity refers to a collection of fossils that, although time averaged to some degree, are believed to have interacted directly or indirectly during life. Paleocommunities

are used for a number of paleontological inferences, from paleoecological interpretations to depth gradient recreation to high-resolution biostratigraphy (Holland and Patzkowsky, 2007; Redman et al., 2007; Forcino et al., 2012a; Schneider et al., 2012). Multivariate statistics enable researchers to visualize and statistically evaluate the variation among paleocommunities. Without multivariate statistical tools, it would be an arduous task to fully characterize the often complex relationships among paleocommunity samples within large paleocommunity datasets.

The most commonly used multivariate technique is ordination (Green, 1979; Clarke, 1993; Legendre and Legendre, 1998; Anderson, 2001; Hendy et al., 2007; Forcino et al., 2010; Jardine et al., 2012; Yasuhara et al., 2012). Ordination is an exploratory multivariate visualization tool that allows researchers to examine the multidimensional relationships among samples in fewer dimensions (Clarke and Ainsworth, 1993; McCune and Grace, 2002). Because of the manner in which ecological data is constructed—a series of samples containing taxonomic objects, each object having a particular abundance value—ordination is the standard way to visualize the similarities and differences among samples. Samples that have more similar taxonomic distributions plot closer together in the ordination space. A researcher can then examine an ordination plot for a pattern or grouping of samples that correlate with independent variables. For example, samples may separate within ordination space based on the lithology of the rocks from which each sample was obtained (Schneider et al., 2012; Forcino et al., in review). However, ordination alone is not a statistical test.

Multivariate goodness-of-fit tests are used for comparing two paleocommunity datasets, for exploring how a dataset of environmental variables compares to a paleocommunity dataset, or for testing for spatial or temporal autocorrelation (Mantel, 1967; Judas et al., 2002; Fall and Olszewski, 2010; Forcino, 2012; Forcino et al., 2012b). Two primary methods for conducting these goodness-of-fit comparisons are the Mantel test (Mantel, 1967; Jackson and Harvey, 1989; Manly, 1997; Dutilleul et al., 2000; Legendre, 2000; Legendre et al., 2005; Legendre and Fortin, 2010) and the Procrustes randomization test (PROTEST; Jackson, 1995; Peres-Neto and Jackson, 2001).

The Mantel test assesses the goodness-of-fit between two multivariate datasets by permuting each of the elements in a calculated matrix of dissimilarity indices (values that quantify the dissimilarity between each object in a taxon-sample matrix) to derive a distribution of correlation values (Mantel, 1967; Clark and Ainsworth, 1993; Legendre and Legendre, 1998; Legendre, 2000; Fall and Olszewski, 2010). This allows for comparisons between a number of types of datasets: (1) two paleocommunity datasets, (2) a paleocommunity dataset and a dataset of environmental variables, or (3) a paleocommunity dataset and an a priori predicted or modeled dataset. The next step then tests whether the distribution of correlation values differs significantly from that expected due to chance. The resulting R-statistic is similar to the Pearson's Product Moment Correlation Coefficient (r); with increasingly similar datasets, the

Mantel R-statistic will approach 1.

The PROTEST, a more recently developed method, is based on the procrustes transformation, which has long been popular in the morphometrics field (Gower, 1975; Bookstein, 1985; Rohlf and Slice, 1990; Chapman, 1990) and has also been employed by ecologists (Gower, 1971). The PROTEST bases its results on data drawn from ordination of the dataset. Ordination, specifically non-metric multidimensional scaling (NMDS), may not always assign the maximum variation in ordination space to the first axis. Moreover, two similar ordinations may appear superficially dissimilar because one ordination may consist of samples that are reflected or rotated compared to the second ordination. To address these possibilities, the first step in PROTEST is to perform a Procrustes transformation, which minimizes the sum-of-squares deviations between the two ordination results through translation, reflection, rotation, and dilation. Thus, the two ordinations are reoriented such that they are aligned as closely as possible in ordination space, which permits a more accurate assessment of similarity. The residuals between the two ordinations post-transformation are calculated and produce an $m^2$-value. Like the R-statistic for the Mantel test, the PROTEST $m^2$-value is similar to the r-value resulting from a Pearson's Product Moment Correlation; the closer $m^2$ is to 1, the more similar the two ordinations. After the Procrustes transformation, PROTEST randomly permutes the ordination scores for all samples for 999 iterations, and an $m^2$-value is calculated for each iteration. A realized p-value, indicating the significance of the $m^2$-value, is then calculated by determining the percentage of iterations in which the $m^2$-values from the randomized iterations are greater than the $m^2$-value of the actual dataset.

One advantage of the PROTEST compared to the Mantel Test is that it can statistically evaluate the goodness-of-fit between two ordination results. This is important because researchers often compare communities by examining the ordinations, not by looking at the raw data. Conversely, there is a point to be made that the underlying data structure, prior to being manipulated by the ordination, is more important to evaluate statistically. Both of these methods are used throughout the research of this thesis.

2. Bibliography

Anderson, M. J. 2001. A new method for non-parametric multivariate analysis of variance. Austral Ecology 26:32–46.

Bambach, R. K. 1993. Seafood through time: changes in biomass, energetics, and productivity in marine ecosystems. Paleobiology 19:372–397

Bambach, R. K, A. H. Knoll, and J. B. Well. 2004. Origination, extinction, and mass depletions of marine diversity. Paleobiology 30:522–542.

Barnosky, A. D., N, Matze, S. Tomiya, G. O. U. Wogan, B. Swart, T. B. Quentel, C. Marshall, J. L. McGuire, E. L. Lindsey, K. C. Maguire, B. Mersey, and E. A. Ferrer. 2011. Has the Earth's sixth mass extinction already arrived? Nature 471: 51–57.

Bennington, J.B. 2003. Transcending patchiness in the comparative analysis paleocommunities: A test case from the Upper Cretaceous of New Jersey. PALAIOS 18:22–33.

Bonelli Jr., J. R., C. E. Brett, A. I. Miller, and J. B. Bennington. 2006. Testing for faunal stability across a regional biotic transition: quantifying stasis and variation among recurring coral-rich biofacies in the Middle Devonian Appalachian Basin. Paleobiology 32:20–37.

Bookstein, F., B. Chernoff, R. Elder, J. Humphries, G. Smith, R. and Strauss. 1985. Morphometrics in Evolutionary Biology. The Academy of Natural Sciences of Philadelphia, Philadelphia, Pennsylvania.

Ceballos, G., and P. R. Ehrlich. 2002. Mammal population losses and the extinction crisis. Science 296:904–907.

Chapman, R. E. 1990. Conventional procrustes approaches. Pp. 251–278 *in* F. J. Rohlf, and F. L. Bookstein, eds. Proceedings of the Michigan Morphometrics Workshop, Special Publication Number 2. University of Michigan Museum of Zoology, Ann Arbor, Michigan.

Clapham, M. E., D. J. Bottjer, C. M. Powers, N. Bonuso, M. L. Fraiser, P. J. Marenco, S. Q. Dornbos, and S. B. Pruss. 2006). Assessing the ecological dominance of Phanerozoic marine invertebrates. PALAIOS 21:431–441.

Clapham, M. E., and N. P. James. 2008. Paleoecology of Early-Middle Permian marine communities in eastern Australia: response to global climate change in the aftermath of the late Paleozoic ice age. PALAIOS 23:738–750.

Clapham, M. E., and D. J. Bottjer. 2007. Permian marine paleoecology and its implications for large-scale decoupling of brachiopod and bivalve abundance and diversity during the Lopingian (Late Permian). Palaeogeography, Palaeoclimatology, Palaeoecology 249:283–301.

Clarke, K. R. 1993. Non-parametric multivariate analysis of changes in community structure. Austral Ecology 18:117–143.

Clarke, K. R., and M. Ainsworth. 1993. A method and linking multivariate community structure to environmental variables. Marine Ecology 92:205–219.

Dirzo, R., and Raven, P. H. 2003. Global state of biodiversity and loss. Annual Review of Environment and Resources 28:137–167.

Dutilleul, P., J. D. Stockwell, D. Frigon, and P. Legendre. 2000. The Mantel Test versus Pearson's correlation analysis: assessment of the differences for biological and environmental studies. Journal of Agricultural, Biological, and Environmental Statistics 5:131–150.

Fall, L. M., and T. D. Olszewski. 2010. Environmental disruptions influence taxonomic composition of brachiopod paleocommunities in the Middle Permian Bell Canyon Formation (Delaware Basin, West Texas). PALAIOS 25:247–259.

Forcino, F. L. 2012. Multivariate assessment of the required sample size for paleocommunity research. Palaeogeography, Palaeoclimatology, Palaeoecology 315:134–141.

Forcino, F. L., E. J. Richards, L. R. Leighton, N Chojnacki, and E. S. Stafford. 2012a. The sensitivity of paleocommunity sampling strategy at different spatiotemporal scales. Palaeogeography, Palaeoclimatology, Palaeoecology 313:246–253.

Forcino, F. L., E. S. Stafford, L. R. and Leighton. 2012b. Perception of paleocommunities at different taxonomic levels: how low must you go? Palaeogeography, Palaeoclimatology, Palaeoecology 365:48–56.

Forcino, F. L., E. S. Stafford, J. J. Warner, A. E. Webb, L. R. Leighton, C. L. Schneider, T. S. Michlin, L. M. Palazzolo, J. R. Morrow, and S. A. Schellenberg. 2010. Effects of data categorizations on paleocommunity analysis: a case study from the Pennsylvanian Finis Shale of Texas. PALAIOS 25:144–157.

Gower, J. C. 1971. Statistical methods for comparing for comparing different multivariate analyses of the same data. Pp. 138–149 *in* F. R. Hodson, K. G. Kendell, and P. Tautu, eds. Mathematics in Archaeology and Historical Sciences, Edinburgh University Press, Edinburgh.

Gower, J. C. 1975. Generalized procrustes analysis. Psychometrika 40:33–51.

Green, R. H. 1979. Sampling Design and Statistical Methods for Environmental Biologists. Wiley, New York.

Heim, N. A. 2009. Stability of regional brachiopod diversity structure across the Mississippian/Pennsylvanian boundary. Paleobiology 35:393–412.

Holland, S. M., and M. E. Patzkowsky. 2007. Gradient ecology of a biotic invasion: biofacies of the Type Cincinnatian Series (Upper Ordovician), Cincinnati, Ohio Region, USA. PALAIOS 22:392–407.

Hughes, J. B., G. C. Daily, and P. R. Ehrlich. 1997. Population diversity: its extent and extinction. Science 278:689–692.

Jackson, D. A. 1995. PROTEST: A Procrustean randomization test of community environment concordance. Écoscience 2:297–303.

Jackson, D. A., and H. H. Harvey. 1989. Biogeographical associations in fish assemblages: Local versus regional processes. Ecology 70:1472–1484.

Jardine, P. E., G. J. Harrington, and T. A. Stidham. 2012. Regional-scale spatial heterogeneity in the late Paleocene paratropical forests of the U.S. Gulf Coast. Paleobiology 38:15–39.

Judas, M., K. Dornieden, and U. Strothmann. 2002. Distribution of carabid beetle species at the landscape-level. Journal of Biogeography 29:491–508.

Kowalewski, M., K. Gurs, J. H. Nebelsick, W. Oschmann, W. E. Piller, and A. P. Hoffmeister. 2002. Multivariate hierarchical analyses of Miocene mollusk assemblages of Europe: Paleoceanographic, paleoecological, and biostratigraphic implications. GSA Bulletin 114:239–256.

Kruskal, J. B. 1964. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychomerika 29:1–27.

Legendre, P., and L. Legendre. 1998. Numerical Ecology. Elsevier. Amsterdam, The Netherlands.

Legendre, P. 2000. Comparison of permutation methods for the partial correlation and partial Mantel Tests. Journal of Statistical Computation and Simulation 67:37–73.

Legendre, P. 2005. Species associations: the Kendell Coefficient of Concordance revisited. Journal of Agricultural, Biological, and Environmental Statistics,10:226–245.

Legendre, P., D. Borcard, and P. Peres-Neto. 2005. Analyzing beta diversity: partitioning the spatial variation of community composition data. Ecological Monographs 75:435–450.

Legendre, P., and M. Fortin. 2010. Comparison of the Mantel test and alternative approaches for detecting complex multivariate relationships in the spatial analysis of genetic data. Molecular Ecology 10:831–844.

Leibold M. A., M. Holyoak, N. Mouquet, P. Amaraekera, J. M. Chase, M. F. Hoopes, R. D. Holt, J. B. Shurin, R. Law, D. Tilman, M. Loreau, and A. Gonzales. 2004. The metacommunity concept: a framework for multi-scale community analysis. Ecology Letters 7:601–613.

Mantel, N. 1967. The detection of disease clustering and a generalized regression approach. Cancer Research 27:209–220.

Maurer, B.A. 1999. Untangling Ecological Complexity. The University of Chicago Press, Chicago.

McCune, B., and J. B. Grace. 2002. Analysis of Ecological Communities. MjM Software Design, Gleneden Beach, Oregon.

McGhee, G. R. 1996. The Late Devonian Mass Extinction: The Frasnian/Famennian Crisis. Columbia University Press, New York.

Olszewski, T. D., and M. E. Patzkowsky. 2001. Measuring recurrence of marine biotic gradients: a case study from the Pennsylvanian–Permian Midcontinent. PALAIOS 16:444–460.

Peres-Neto, P. R., and D. A. Jackson. 2001. How well do multivariate datasets match? The advantages of a Procrustean superimposition approach over the Mantel Test. Oecologia 129:169–178.

Raup, D. M., and J. J. Sepkoski. 1982. Mass extinctions in the marine fossil record. Science 215:1501–1503.

Redman, C. L., L. R. Leighton, S. A. Schellenberg, C. N. Gale, J. L. Neilsen, D. L. Dressler, and

M. K. Klinger. 2007. Influence of spatiotemporal scale on the interpretation of paleocommunity structure: lateral variation in the Imperial Formation of California. PALAIOS 22:630–641.

Rohlf, F. J., and D. Slice. 1990. Extensions of the procrustes method for the optimal superimposition of landmarks. Systematic Zoology 39:40–59.

Sepkoski, J. J., R.K. Bambach, D. M. Raup, and J. W. Valentine. 1981. Phanerozoic marine diversity: a strong signal from the fossil record. Nature 293:435–437.

Schneider, C. L., F. L. Forcino, K. Barclay, and L. R. Leighton, L.R. 2012. Significance of paleocommunity analysis for interpreting the ecology and geology of the Devonian Waterways Formation, northeastern Alberta, Canada. Geological Society of America Program with Abstracts 44:625.

Stafford, E. S., and L. R. Leighton. 2011. Vermeij Crushing Analysis: A new old technique for estimating crushing predation in gastropod assemblages. Palaeogeography, Palaeoclimatology, Palaeoecology 305:123–137.

Tyler, C. L., and L. R. Leighton. 2011. Detecting competition in the fossil record: support for character displacement among Ordovician brachiopods. Palaeogeography, Palaeoclimatology, Palaeoecology 307:205–217.

Vinatier, F., P. Tixier, P. Duyck, and F. Lescourret. 2011. Factors and mechanisms explaining spatial heterogeneity: a review of methods for insect populations. Methods in Ecology and Evolution 2:11–22.

# Chapter 2. The sensitivity of paleocommunity sampling strategy at different spatiotemporal scales*

## 1. Introduction

Community paleoecology utilizes complete fossil assemblages to determine the mechanisms of spatiotemporal ecological and environmental variation, aiding in the pursuit of the processes that structure ecosystems and the causes of ecosystem collapse and extinction (Jablonski, 1998; Olszewski and Patzkowsky, 2001; Bonelli et al., 2006; Clapham et al., 2006; Clapham and James, 2008; Heim, 2009). Paleocommunities (herein referring to marine invertebrate fossil assemblages used to infer environmental or ecological gradients) provide a wealth of information regarding taxonomic interactions and environmental tolerances on local scales and how these processes scale up to regional, continental, and global scales (Bambach, 1993; Kowalewski et al., 2002; Bambach et al., 2004; Clapham and Bottjer, 2007).

Because modern ecological research is limited to human time scales (i.e., years to thousands of years), modern ecologists usually examine community relationships through space (Downes et al., 1993; Rosenzweig, 1995; Bustamante and Branche, 1996; Boström and Bonsdorff, 1997; Broitman et al., 2001; Harte et al., 2005**;** Hereu et al., 2008). A primary difference between paleoecological studies and modern ecological studies is that paleoecological studies have the added dimension of deep geologic time, which enables examinations of ecological persistence, turnover, and extinction effects through time (Pandolfi, 1996; Boyer et al., 2004; Clapham and James, 2008; Heim, 2009; Layou, 2009; Olszewski and Erwin, 2009). However, having to take into account both temporal and spatial dimensions also leads to complications when determining the most appropriate sampling and analytical methods required for conducting paleocommunity research.

Forcino et al. (2010) examined marine invertebrate paleocommunity variation through a 5 meter stratigraphic section of the Virgilian (Gzhelian) Finis Shale of Texas and identified two distinct paleocommunities, one occupying the lower portions of the section (older) and one the upper (younger). Although this temporal change in paleocommunity structure was clear at this one stratigraphic section, this result does not provide any evidence for the distribution of communities over broader spatial scales of the Finis Shale. Thus, additional sampling of the Finis Shale laterally would add the supplementary dimension that might lead to a different result.

Recent studies examining the spatial sampling procedures required for studying paleocommunities have demonstrated that smaller, replicate (within bed or unit) samples produce more robust community patterns than one large bulk sample (Lafferty et al., 1994; Bennington, 2003; Webber, 2005; Zuschin et al., 2006; Zambito et al., 2008). The argument of the above

---

* *A version of this chapter has been published. Forcino, F.L., Richards, E.J, Leighton, L.R., Chojnacki, N. and Stafford, E.S. (**2012**) The sensitivity of paleocommunity sampling strategy at different spatiotemporal scales. Palaeogeography, Palaeoclimatology, Palaeoecology 313:246-253.*

studies is that large, single samples do not recover the complete diversity of a paleocommunity because such samples fail to capture variation caused by faunal patchiness and the sparse distribution of rare taxa. Based on the results of these studies, Forcino et al. (2010) may not have fully captured the structure of the Finis Shale paleocommunities because only one stratigraphic section was examined.

Here, to capture possible community changes through time in the Finis Shale more completely, I explore the lateral variation of the Finis Shale communities. Because the Finis Shale outcrops at easily accessible locations that also occur along strike at approximately regular intervals, the exposures provide an opportunity to examine how, and if, the temporal change in paleocommunities within the Finis Shale varied through space. For the present study, I collected 29 samples from six localities of the Finis Shale and conducted multivariate paleocommunity analyses in order to determine if additional lateral sampling leads to a change in the paleocommunity signal (the information and patterns produced based on the taxonomic distribution within and among samples) determined by Forcino et al. (2010). A difference in paleocommunity signal would be evidence of the importance of spatial sampling; any ecological conclusions based on only the original section would be based on incomplete data. In contrast, a consistent and laterally persistent paleocommunity signal would serve as evidence that complete spatial sampling may not always be necessary for all paleocommunity studies. However, such a result would not be very informative, as it would not be clear to a researcher whether lateral sampling was necessary *a priori*. Thus, it would be of additional interest to determine under what circumstances extensive spatial sampling might no longer be essential. One possibility may be the scale of the study, especially the temporal extent of sampling.

Paleocommunity studies examining faunal persistence, coordinated stasis, and high-resolution environmental variables (e.g., within-stratigraphic-unit depth gradient) often are examining communities within only a few stratigraphic horizons, beds, or units (Zambito et al., 2008). Often such studies limit themselves to one stratigraphic formation. The influence of the processes operating at these finer scales is inherently different (or at least less averaged by spatiotemporal variables) than those at larger scales (multiple stratigraphic formations). At small spatial scales, fine-scale controls on community composition may include microhabitat or biotic interactions, while at coarser scales water depth is often perceived as a primary controlling factor because it takes into account many other oceanographic parameters (Holland, 2005; Redman et al., 2007). Therefore, the types of ecological explorations and hypotheses examined are often quite different between smaller and larger scales.

Communities ordinated along a gradient reflect changes in taxonomic composition driven by some environmental variable or variables. Although the variables driving the gradient may be unknown (and cannot be identified based on the ordination alone without independent

data), the processes underlying the gradient may operate on different scales, such that a change in sampling or analytical scale would alter the paleocommunity signal. For example, would communities sampled at a greater temporal scale be more likely to reveal a gradient indicative of a change in environment through time, rather than through space? In other words, would the spatial signal be obscured at greater temporal scales? If so, then there would be some temporal scale at which spatial variation within a single point in time would be reduced to noise in comparison to the overriding signal through time. As the choice of operational scale is dictated by the question of interest, scale considerations of this sort could affect decisions regarding sampling strategy.

Because Forcino et al. (2010) found two distinct paleocommunity sets separated by the approximate mid-point of the stratigraphic extent of the Finis Shale, there is the potential for a test of the hypothesis that the importance of lateral sampling may vary with the temporal scale under study. The stratigraphic section can be divided into two halves, and the multivariate analysis repeated only on one half. This second analysis would examine how changing the temporal scale (reducing it to approximately one half of the previous analysis) might affect any signal derived from lateral variation.

## 2. Methods

### 2.1 Geologic Background

The Finis Shale was deposited along the paleoequator on the Eastern Shelf of the Midland Basin in what is now Texas, USA (Figure 2-1). During the Virgilian (Gzhelian), tectonic activity was occurring paleonorth, paleoeast, and paleosouth of the Midland Basin (Brown, 1973; Heckel, 1977). The Amarillo-Wichita Uplift separated the Midcontinent and the Midland Basin and was active during the Late Pennsylvanian, leading to massive amounts of fine siliciclastic input into the smaller Midland basin (Algeo and Heckel, 2008). However, the Finis Shale lacked coarse siliciclastic input; rather, terrigenous mud settled out of suspension in a calm, low-energy environment over approximately two million years (Cheney, 1940; Brown, 1973). As a result, sediment grain size remains relatively constant, both laterally (> 10 km) and vertically (> 5 m), within the Finis Shale. Finis Shale strata are essentially flat-lying over great distances, with only local variations in dip attributed to either syndepositional topographic variation or post-burial sediment compaction. Within the Finis Shale, I chose outcrops on a transect paralleling the shoreline of the shallow Eastern Shelf. Thus, I am able to compare spatial and temporal variation at multiple scales within a system in which sedimentary regime and water depth is a relatively controlled factor.

Each stratigraphic section sampled ranged in vertical extent from ~4 m to 6 m. The base of each stratigraphic section was the lowest exposed point of the Finis Shale. The contact between the Finis Shale and the underlying Homecreek Limestone was not exposed at any of the sampled stratigraphic sections. At the Jacksboro Spillway locality, the lower bound of the three

stratigraphic sections sampled was a sandstone lens within the Finis Shale. The upper contact at each stratigraphic section between the Finis Shale and the Jacksboro Limestone was highly weathered. The limestone often slumped down into portions of the Finis Shale. Due to these two factors, samples are identified stratigraphically by their position in meters above the base of each outcrop. Each stratigraphic section was ~5 m in horizontal height. For additional information about the Finis Shale see Forcino et al. (2010).

2.2 Sampling

Samples were collected from four localities of the Finis Shale in North Central Texas (Figure 2-1; Table 2-1; Schindel et al., 1982; Grossman et al., 1991). The four localities were selected along a ~10 km northeast-southwest transect of the Finis Shale. Within the extensive Jacksboro Spillway locality, I sampled three vertical sections (West, Middle, and East Jacksboro Spillway) along outcrop to quantify within-outcrop (< 1 km) lateral variation (Figure 2-1), for a total of six sampled stratigraphic sections.

At each of the sampled stratigraphic section, ~4 L (one gallon Ziploc bag) bulk sediment samples were collected vertically every ~0.5 to 1 meter in an attempt to capture the full range of temporal variability that may exist at each locality. Shale units were bulk sampled in order to obtain representative abundances of the communities.

A sample size of 4 L was chosen because a pilot study demonstrated that at the Finis Shale Spillway West locality, 2 L of bulk sediment per sample produced the same ordination-based community signal (the information and patterns revealed based on the taxonomic distribution through and among samples) as a 13 L bulk sediment sample (Forcino and Leighton, 2010a). In addition, Forcino (in press)—in which two Finis Shale paleocommunity datasets were included as a portion of the study—found that a sample size of 25 to 50 specimens is sufficient for multivariate analyses of paleocommunities. Based on these results, 2 L samples would be sufficient for the aims of the present study, but to be conservative, 4 L samples were collected.

Samples were soaked in water with a mild detergent for one to seven days to disaggregate the sediment. Samples for which detergent was ineffective were soaked in a three percent hydrogen peroxide solution for 24 hours. If water and detergent were ineffective, samples were sonicated for 30 to 60 minutes. All samples were wet-sieved and the fossil specimens were sorted into large- (> 2.8 mm) and small- (< 2.8 mm) size-fractions. Only the > 2.8 mm size-fraction was considered in subsequent analyses in order to simulate the collection protocols of most field-collection-based macro-invertebrate paleoecological community studies; this is also consistent with the protocol of Forcino et al., (2010). All brachiopod, mollusk, echinoid, bryozoan and coral specimens were sorted and identified to genera. Because I found only disarticulated columnals of crinoids (no parts of the cup were in any samples), crinoids were differentiated based on the morphology of the columnal, and so represent morphotaxa.

Abundance counts were completed using the minimum number of individuals (MNI) technique described by Gilinsky and Bennington (1994). This technique gives a count of one to all articulated brachiopod or bivalve specimens and a count of one to the most abundant valve (either the pedicle or brachial and either the left or right) of all disarticulated specimens. For a brachiopod the cardinal process had to be present on the brachial value or the umbo or pedicle foramen had to be present on the pedicle value. For a bivalve specimen to be counted, the umbo had to be present. However, if there was any identifiable fragment of a single specimen of a single taxon within a sample, even if there was not a complete specimen in that sample, an abundance count of one was assigned to that taxon for that sample. For example, if a given sample contained only fragments of the bivalve *Astartella* without any evidence of the umbo, then this genus was still given an abundance count of one; the taxon does exist in the sample and counting it as absent would be incorrect. For colonial taxa (bryozoans and tabulate corals) and easily disarticulated taxa (crinoids), one individual was counted for every 1 cm of length using the maximum length of each fossil fragment. This technique has been used in previous studies (Patzkowsky and Holland, 1999; Holland and Patzkowsky, 2004) as a means of equating the abundances of colonial and easily disarticulated taxa to those of solitary taxa (brachiopods and mollusks). A length of 1 cm was used because that was approximately the mean length of solitary individuals in most samples of the Finis Shale. Forcino and Leighton (2010b) found that, in the Finis Shale, the paleocommunity signal produced by the taxon-sample matrix using the technique of equating 1 cm of length to one individual solitary specimen is significantly correlated with the paleocommunity signal obtained when using presence or absence of colonial and easily disarticulated taxa in the taxon-sample matrix ($p < 0.001$).

Paleocommunity variation was examined using ordination and analysis of similarity (ANOSIM). Ordination analyses were performed using the vegan package (Oksanen et al., 2010) in R 2.11.1 (R Development Core Team, 2010). Ordination analysis is an effective means of expressing multidimensional relationships among objects by identifying the primary variation in the dataset and expressing that variation within a few axes. The resulting ordination plot orders the objects along axes according to their similarity, with more similar objects plotting closer in space (Beals, 1984; Ludwig and Reynolds, 1988). In the present study, objects are samples ordinated based on the abundances of their constituent taxa using Non-metric multidimensional scaling (NMDS) ordination. NMDS is a popular ordination technique in both ecology (Minchin, 1986; Clarke and Ainsworth, 1993; Laughlin and Abella, 2007) and paleoecology (Bonelli et al., 2006; Clapham and James, 2008; Layou, 2009). Several authors have argued that NMDS is the most effective ordination technique for capturing community gradients (Minchin, 1986; McCune and Grace, 2002; Bush and Brame, 2010).

Ordinations were performed using the Sorenson (Bray-Curtis) dissimilarity measurement between samples. Because bulk-sediment shale samples contained the same

sediment volume, the analyses did not require any standardizations or percent transformations. Therefore, the Sorenson dissimilarity measurement was used to retain absolute abundance information. Comparing absolute abundances may retain more ecological information (e.g., similarities and differences among communities) about the once-living community (Finnegan and Droser, 2005) and should be utilized when comparing communities collected with a consistent sample size and protocol (Clarke and Warwick, 1994). In contrast, when multiple lithologies or sample sizes are examined, abundances are often converted to percentages or presence/absence of taxa (Clarke and Warwick, 1994; Zuschin et al., 2006).

2.3 Statistical Analysis

In addition to analysis of the dataset containing all 29 samples collected from the Finis Shale, a subset of 19 samples from the lower half of each stratigraphic section (the stratigraphically older samples) was ordinated separately. I wanted to compare this finer temporal scale of the subset to the coarser temporal scale of the full data set to determine if a change in scale caused the paleocommunity signal to change. I did not examine the upper stratigraphic samples alone because there were only 10 samples. Conducting a multivariate analysis using only 10 samples may not yield statistically strong results. If paleocommunity signals of the complete 29 samples demonstrated a gradient or pattern through time, while the paleocommunity signal of just the lower-stratigraphic-half samples demonstrated a gradient or pattern through space, this would be evidence to support our hypothesis that the lateral extent of sampling required for paleocommunity research is sensitive to the scale under study. However, if the paleocommunity signal between both the complete set of 29 samples and the subset of the lower-stratigraphic-half samples demonstrates a gradient or pattern through just time or just space, then this would be evidence against our hypothesis.

Sample localities and relative stratigraphic position were mapped onto the ordination results and examined for any groupings (clouds) based on either spatial or temporal variation (Figures 2-2 and 2-3). Locality-based clouds or gradients would indicate a strong spatial control on the distribution of communities; the primary environmental variable driving the ordination, whatever that variable happens to be, is changing through space more so than through time. If clouds or gradients are based on a non-geographic variable (i.e., stratigraphic position), a temporal variable (a variable that is changing mostly through time rather than space) may be controlling the distribution of communities. The other possible outcome may be a lack of data clouds or gradients in ordination space, which would provide no evidence for a strong spatial or temporal signal.

Using the Bray-Curtis dissimilarity matrix, analysis of similarity (ANOSIM) was used to test for significant differences between localities and stratigraphic units (Clarke and Warwick, 1994). ANOSIM is a non-parametric multivariate form of an analysis of variance that tests for differences between groups of dissimilarity values. ANOSIM converts the Bray-Curtis

dissimilarity values into ranks and compares the distributions of the ranks of various groups of dissimilarity values by permuting the dissimilarity values. The groups are significantly different (greater r-values) if there is limited overlap in the rankings between dissimilarity values.

One set of spatial samples was tested using ANOSIM: the six sample localities were used as six different groups of samples. Two different sets of stratigraphic samples were also tested using ANOSIM: (1) the upper stratigraphic samples versus the lower stratigraphic samples, and (2) the relative stratigraphic position of samples. Set (1) consists of two groups of samples, a lower and upper stratigraphic grouping. The division line of the two groups was the vertical midway point of each outcrop. Set (2) consists of seven groups of samples defined by their relative stratigraphic position (Table 2-2).

3. Results

The coarser-scale taxon-sample abundance matrix consisted of 29 samples totaling 5143 specimens from 70 genera with a mean sample size of 177 specimens, a minimum sample size of 36 specimens, and a maximum samples size of 404 specimens. These samples separated into two clouds in the NMDS ordination (Figure 2-2). The two clouds separated at approximately NMDS axis one scores of zero. Using ANOSIM, there is a significant difference between the two stratigraphic groupings (upper and lower half) of samples ($R = 0.56$, $p < 0.001$; Table 2-2) and a significant difference between at least one of the seven relative stratigraphic position groupings ($R = 0.26$, $p = 0.005$). However, there is no significant difference between the six locality-based groupings ($R = 0.06$, $p = 0.208$). Within the cluster containing the lower-stratigraphic samples (NMDS axis-one score less than 0), these 19 samples display a spatial pattern. The samples from the southeastern two localities (Pankey Property and Causeway Road) plot at the highest NMDS axis-two scores of the cluster. The samples from the Jacksboro Spillway plot at slightly lower NMDS axis-two scores. The samples from Cemetery Road plot at the lowest NMDS axis-two scores of this cluster (Figure 2-2A).

The 19 lower-stratigraphic samples were ordinated to examine the community signal on the finer temporal scale (Figure 2-3). This dataset contained samples from the lower half of each stratigraphic section. This dataset consisted of 2947 specimens from 55 genera with a mean sample size of 155 specimens, a minimum sample size of 36 specimens, and a maximum sample size of 404 specimens. These samples separate based on sample location along the northeast-southwest transect. Samples from the Cemetery Road, Spillway, and the southwestern (the Causeway and Pankey localities combined) localities form distinct clusters in the NMDS ordination. Using ANOSIM, there is a significant difference between at least one of the six locality-based groupings of lower stratigraphic samples ($R = 0.29$, $p = 0.007$; Table 2-2). There is no significant difference between the seven relative stratigraphic position groupings ($R = 0.01$, $p = 0.43$).

4. Discussion

4.1 Lateral Variation of the Finis Shale paleocommunities

Forcino et al. (2010) examined paleocommunity variation through the stratigraphic section at the Spillway West locality. Their 13 samples produced the same general separation in ordination space as the 29 samples from the present study. The samples from the lower portions of the section (~the lower half) were dominated by the brachiopod *Crurithyris*, and plotted along lower ordination axis-one scores. The samples from the upper portions of the section (~the upper half) were dominated by the brachiopod *Rhipidomella*, and plotted along higher ordination axis-one scores. The two sets of communities identified in Forcino et al. (2010) were laterally persistent, and the two communities were separated in time. Multivariate results demonstrate strong and consistent separation between these communities (Figure 2-2).

In the present study, extensive lateral sampling at the coarser temporal interval produced the same community separation at any single section as any other section and as found by Forcino et al. (2010). There was a separation of samples in the lower-stratigraphic cluster—the cluster along lower NMDS one-one scores—based on location; this is the same separation demonstrated when those samples are ordinated separately (Figure 2-3). However, this minor spatial pattern is still obscured by the strong temporal signal (Table 2-2; Figure 2-2). This consistent community signal with the addition of lateral sampling, both within and among outcrops, appears to contradict earlier research that suggests multiple lateral samples are required within any stratigraphic unit in order to precisely capture the paleocommunity variation (Lafferty et al., 1994; Bennington, 2003; Webber, 2005; Bonelli et al., 2006; Zuschin et al., 2006; Zambito et al., 2008; Hendy et al.,. 2009). So, why is there a discrepancy between the bulk of previous published research and the present study? One hypothesis is that the difference in temporal scale (specifically the extent of the sampling through time) between this study and previous studies caused the difference; the temporal scale influences the relative importance of spatial signal.

4.2 Sensitivity of the Finis Shale paleocommunity signal depending on scale

Because the two distinct Finis Shale communities (older-*Crurithyris*-dominated and younger-*Rhipidomella*-dominated) found by Forcino et al. (2010) consistently separate at the half-way point of each of the stratigraphic sections sampled, this provided the means to test our hypothesis that the importance of lateral sampling for paleocommunity research may vary with the temporal scale under study. After dividing the stratigraphic section in two halves, I examined the paleocommunity signal of the samples collected from just the lower half of the Finis Shale at all six sections and compared this to the paleocommunity signal of all 29 samples from the complete stratigraphic section of the Finis Shale.

The primary controls of the Finis Shale paleocommunities signal vary between the analyses of the two different stratigraphic scales. At the finer temporal scale containing 19 samples from the lower-*Crurithyris*-dominated half of each stratigraphic interval, the community signal varies along a gradient through space. Evidence for the strong spatial signal at the fine-

scale includes (1) sample separation in ordination space based on the sample's location (from the lower left to the top right of the ordination, there is a gradient representing change from the northeast to the southwest; Figure 2-3) and (2) a significant difference between locality-based groupings using ANOSIM, with no significant difference between stratigraphic-based groupings (Table 2-2).

The 29 Finis Shale samples from the complete stratigraphic section of each of the six outcrop locations represent a larger temporal scale. These communities vary from older-*Crurithyris*-dominated samples to younger-*Rhipidomella*-dominated samples and a few coral- and bryozoan-dominated samples. With the addition of this one set of 12 samples from an additional ~2.5 m of stratigraphic section, there is a change from a spatial factor to a temporal factor controlling the primary distribution of communities. This is not to suggest that whatever unknown environmental variable influencing the distribution of lower Finis Shale paleocommunities is not operational at the larger temporal scale. However, when additional samples through time are included in the analysis, the paleocommunity signal from this unknown environmental variable is overwhelmed by another variable that changes through time. The samples with lower axis-one scores are all from lower stratigraphic portions of each outcrop, while the samples with higher axis-one scores are from upper stratigraphic portions of each outcrop, with a clear division of sample clouds (Figure 2-2). Although the minor spatial gradient is visually evident in the ordination of the complete 29 sample dataset, the pattern is much more difficult to discern than the primary temporal pattern. In addition, no statistically significant pattern was found through space (Table 2-2). The minor variation in the temporal scale of the study, leading to a major change in the potential interpretation of a primary factor controlling the community distribution in ordination space, is evidence that the results of a paleocommunity study are extremely sensitive to the temporal extent of sampling.

If examining one outcrop alone (e.g., Spillway Middle), the stratigraphic ordering of samples is the same (Figure 2-4). The same temporal pattern can be discovered even with very limited lateral sampling. The three lower stratigraphic samples all group together closely followed by sample 4 from the Spillway Middle. Samples 5 and 6 are clearly different and plot much higher along axis two. This demonstrates the strong stratigraphic signal even with only one outcrop's samples.

The results here do not necessarily disagree with those of previous studies arguing for extensive lateral sampling (Lafferty et al., 1994, Bonelli et al., 2006; Zuschin et al., 2006; Zambito et al., 2008). These studies demonstrated that on fine temporal scales, lateral sampling is essential. However, if the temporal extent of these studies were expanded, fewer lateral samples may be required.

Lafferty et al. (1994), Bonelli et al. (2006), and Zambito et al. (2008) examined spatial variation of marine invertebrate communities of the Devonian Hamilton Group of New York.

Lafferty et al. (1994) demonstrated a strong spatial trend among temporally equivalent communities. They found greater variation in communities between localities than laterally equivalent samples within a single locality. Zambito et al. (2008) discovered both fine-scale (within outcrop) and regional scale (50 km to 100 km) community variation along one stratigraphic horizon. Bonelli e al. (2006) also reported significant spatial community variation among local and regional scales. Although there are some minor methodological differences between these three studies, they all similarly discovered within-locality lateral variation in communities and concluded that paleoecological studies must collect several bulk sediment samples per fossil horizon in order to completely describe the community variation.

Zambito et al. (2008) presented ordination-based results that clustered based on the collection location of samples along one stratigraphic horizon, a result similar to the Finis Shale results when only one temporal interval, the lower samples of the Finis Shale, were analyzed (Figure 2-3). However, as Zambito et al. (2008) did not examine samples from different horizons, it was impossible to determine the importance of the temporal signal relative to the spatial signal. Therefore, the conclusions of Zambito et al. (2008) may be specific to studies at a higher spatial resolution, and smaller temporal scale, than the present study.

Lafferty et al. (1994) used polar ordination to examine differences in community composition through ~200 km. Although the samples from each horizon strongly clustered in ordination space based on locality, spatial variation in each of the two stratigraphic horizons was analyzed separately. Lafferty et al. (1994) did not combine all samples collected from the two beds into a single ordination to analyze temporal variation. Such ordinations may reveal similar results to the present study, in which temporal variation obscures the presence of any spatial pattern.

Although this was not the goal of their studies, the data of Bonelli et al. (2006) support the hypothesis that the relative influence of temporal and spatial community signals are scale dependent. Bonelli et al. (2006) examined two stratigraphic horizons from six localities in New York and three in Pennsylvania. Their results are similar to those of the present study. The samples from Bonelli et al. (2006) clustered based on spatial distribution only when each horizon was ordinated separately. However, when their ordination was scaled-up to include both stratigraphic horizons, the samples displayed equal or greater separation based on a temporal, rather than a spatial, trend.

Zuschin et al. (2006) also determined that multiple lateral samples are required per fossil horizon to capture the full richness and evenness of a unit in the middle Miocene Grund Formation of Austria. Furthermore, they also acknowledged the scale- and time-dependent nature of their conclusions. Although they explored community variation using NMDS ordinations, their emphasis was placed on diversity metrics. Their ordinations resulted in similar patterns to

that of Bonelli et al. (2006) and the present study; Zuschin and others' (2006) five stratigraphic horizons formed three stratigraphic-driven clusters in ordination space.

The consistency between the present study and the studies discussed above may also be due to the fact that all use multivariate analytical techniques (e.g., ordination, ANOSIM). Although the complete diversity may not be obtained from fewer samples per stratigraphic horizon, multivariate analytical methods of quantifying community variation may not require the full taxonomic diversity to capture the relative similarities and order of the communities (Forcino et al., 2010; Forcino, in press). Thus, if the intent of the study is primarily to capture an environmental gradient through time using multivariate analytical techniques, then concerns about spatial heterogeneity and a complete sampling of diversity may be unwarranted.

There is certainly information to be gained from sampling the complete lateral extent of the Finis Shale, or any fossiliferous unit, in outcrop (e.g., the complete fossil diversity of the Finis Shale); however, if the purpose of the study were to determine the primary factor controlling any gradients or ordination results, sampling just one outcrop would have been sufficient for an accurate analysis of the Finis Shale through time. Thus, the case study presented here raises the possibility that fewer lateral-equivalent samples are required per stratigraphic bed when examining coarser-scale community patterns.

This is not a comprehensive determination of how many samples are required per horizon, bed, unit, or formation. It is evidence that the current standard practice of paleocommunity researchers can possibly be modified when examining paleoecological variation over larger time scales. This may aid in research that requires a great number of samples over a larger temporal and spatial extent. However, further research is required to determine specific numbers of samples required per horizon, bed, unit, or formation based on the particular scale of interest in any particular study.

## 5. Conclusion

I found no difference between the Finis Shale paleocommunity signals from a single section (Forcino et al. 2010) and those signals obtained by additional lateral sampling across 10 km. This is in contrast to previous research that found that multiple lateral samples per stratigraphic horizon, bed, or unit are required to obtain a precise paleocommunity signal. The primary cause for this difference is the difference in temporal scale between the present study and the previous research, specifically the increase in the temporal extent of sampling. At the finer temporal scale containing 19 samples from the lower half of each stratigraphic interval, the primary pattern of community distribution reveals a spatial pattern. At the larger scale containing 29 samples from the complete stratigraphic extent at six localities of the Finis Shale, a pattern of community distribution through time is manifested and appears to overwhelm any spatial signal. This minor variation in temporal scale, leading to a major change in the primary factor driving the pattern in ordination space, is evidence that patterns revealed at different scales are extremely

sensitive to the temporal extent of sampling. In addition, reevaluation of multivariate data from Lafferty et al. (1994), Zuschin et al. (2006), and Zambito et al. (2008) confirms that variation in temporal scale can cause variation in the paleocommunity signal.

I am not suggesting that the use of multiple lateral samples is in any way detrimental to paleocommunity studies. Our results provide evidence that increasing the number of lateral samples per stratigraphic unit does not increase the accuracy of results when larger temporal scales are studied using multivariate methods. Thus, when examining coarser-scale community variation, sampling effort is better-spent collecting samples from a greater number of stratigraphic units rather than replicating samples laterally.

Although further work and additional case studies on this hypothesis are definitely in order, this serves as evidence that there is some temporal scale at which the temporal community signal will overwhelm that of the spatial scale. Thus, since the question of interest dictates the temporal and spatial scale of the study, if the question requires a broader temporal scale, the problem of lateral variation is probably not as serious as has been thought; community paleoecology examinations at larger scales require fewer samples per stratigraphic horizon, bed, or unit.

## 5. Acknowledgements

The paper is dedicated to the memory of Jared Morrow. His aid in the field was critical to the completion of the project, and his mentoring was essential to the scientific development of the lead author. This research was supported by NSF-EAR-0746072 Grant to Leighton, Stephen Schellenberg, Jared Morrow, and Chris Schneider. Thank you for the help provided by Chris Schneider for locality selection, collection of samples, and feedback on the manuscript. I would like to thank the city of Jacksboro, Texas, for permission to collect samples. I also thank Stephen Schellenberg and Brian Pierce for their assistance in the field, Mark Labbe and David Chesterman for their support in setting up sediment processing facilities at the University of Alberta, and Brian Chatterton for his feedback on the manuscript. Thank you to Andy Bush and Matthew Clapham for helpful reviews of this manuscript.

## 6. Bibliography

Algeo, T.J. and Heckel, P.H., 2008. The Late Pennsylvanian Midcontinent Sea of North America: a review. Palaeogeography, Palaeoclimatology, Palaeoecology. 268, 205-221.

Bambach, R.K., 1993. Seafood through time: changes in biomass, energetics, and productivity in marine ecosystems. Paleobiology. 19, 372-397.

Bambach, R.K, Knoll, A.H. and Well, J.B., 2004. Origination, extinction, and mass depletions of marine diversity. Paleobiology. 30, 522-542.

Beals, E.W., 1984. Bray-Curtis Ordination: an effective strategy for analysis of multivariate ecological data. Advances in Ecological Research, 14, 1-55.

Bennington, J.B., 2003. Transcending patchiness in comparative analysis of paleocommunities: a

test case from the Upper Cretaceous of New Jersey. PALAIOS. 18, 22–33.

Bonelli Jr., J.R., Brett, C.E., Miller, A.I. and Bennington, J.B., 2006. Testing for faunal stability across a regional biotic transition: quantifying stasis and variation among recurring coral-rich biofacies in the Middle Devonian Appalachian Basin. Paleobiology. 32, 20–37.

Boström, C. and Bonsdorff, E. (1997) Community structure and spatial variation of benthic invertebrates associated with *Zostera marina* (L.) beds in the northern Baltic Sea. Journal of Sea Research, 37, 153-166.

Boyer, D.I., Bottjer, D.J. and Droser, M.L., 2004. Ecological signature of Lower Triassic shell beds of the western United States. PALAIOS. 19, 372-380.

Broitman, B.R., Navarrete, S.A., Smith, F. and Gaines, S.D., 2001. Geographic variation of southeastern Pacific intertidal communities. Marine Ecology Progress Series. 224, 21-34.

Brown, L.F., Jr., 1973. Cisco depositional systems in North-Central Texas. Pennsylvanian Depositional Systems in North-Central Texas. Austin, Texas. Bureau of Economic Geology. 57-122.

Bush, A.M. and Brame, R.I., 2010. Multiple paleoecological controls on the composition of marine fossil assemblages from the Frasnian (Late Devonian) of Virginia, with a comparison of ordination methods. Paleobiology. 36, 573-591.

Bustamante, R.H. and Branch, G.M., 1996. Large scale patterns and trophic structure of Southern African Rocky shores: the roles of geographic variation and wave exposure. Journal of Biogeography. 23, 339-351.

Cheney, M.G., 1940. Geology of North-Central Texas. Bulletin of the American Association of Petroleum Geologists. 24, 65-118.

Clapham, M.E. and Bottjer, D.J., 2007. Permian marine paleoecology and its implications for large-scale decoupling of brachiopod and bivalve abundance and diversity during the Lopingian (Late Permian). Palaeogeography, Palaeoclimatology, Palaeoecology. 249, 283-301.

Clapham, M.E., Bottjer, D.J., Powers, C.M., Bonuso, N., Fraiser, M.L, Marenco, P.J., Dornbos, S.Q. and Pruss, S.B., 2006. Assessing the ecological dominance of Phanerozoic marine invertebrates. PALAIOS. 21, 431-4411.

Clapham, M.E. and James, N.P., 2008. Paleoecology of Early-Middle Permian marine communities in eastern Australia: response to global climate change in the aftermath of the late Paleozoic ice age. PALAIOS. 23, 738-750.

Clarke, K.R. and Ainsworth, M., 1993. A method and linking multivariate community structure to environmental variables. Marine Ecology. 92, 205-219.

Clarke, K.R. and Warwick, R.M., 1994. Similarity-based testing for community pattern: the two-

way layout with no replication. Marine Biology. 18, 167-176.

Downes, B.J., Lake, P.S. and Schreiber, E.S.G., 1993. Spatial variation in the distribution of stream invertebrates: implications of patchiness for models of community organization. Freshwater Biology. 30, 119-132.

Finnegan, S. and Droser, M.L., 2005. Relative and absolute abundance of trilobites and rhynchonelliform brachiopods across the Lower/Middle Ordovician boundary, eastern Basin and Range. Paleobiology. 21, 480-502.

Forcino, F.L., in press. Multivariate assessment of the required sample size for community paleoecological research. Palaeogeography, Palaeoclimatology, Palaeoecology.

Forcino, F.L. and Leighton, L.R., 2010a. Determining the smallest sample size needed for paleocommunity research: A case study from the Finis Shale of Texas. Geological Society of America Abstracts with Programs. 42, 72.

Forcino, F.L. and Leighton, L.R., 2010b. Determining the most robust method of measuring disaggregated fossil material: the colonial conundrum. Geological Society of America Abstracts with Programs. 42, 257.

Forcino, F.L., Stafford, E.S., Warner, J.J., Webb, A.E., Leighton, L.L., Schneider, C.L., Michlin, T.S., Palazzolo, L.M., Morrow, J.R. and Schellenberg. S.A., 2010. Effects of data categorizations on paleocommunity analysis: a case study from the Pennsylvanian Finis Shale of Texas. PALAIOS. 25, 144-157.

Gilinsky, N.L. and Bennington, J.B., 1994. Estimating numbers of whole individuals from collections of body parts: a taphonomic limitation of the paleontological record. Paleobiology, 20, 245-258.

Grossman, E.L., Zhang, C. and Yancey, T.E., 1991. Stable isotope stratigraphy of brachiopods from Pennsylvanian shales in Texas. Geological Society of America Bulletin. 103, 953-965.

Harte, J., Conlisk, E., Ostling, A., Green, J.L. and Smith, A.B., 2005. A theory of spatial structure in ecological communities at multiple spatial scales. Ecological Monograph. 75, 179-197.

Heckel, P.H., 1977. Origin of phosphatic black shale facies in Pennsylvanian cyclothems of Midcontinent North America. American Association of Petroleum Geologists Bulletin. 61, 1045-1068.

Heim, N.A., 2009. Stability of regional brachiopod diversity structure across the Mississippian/Pennsylvanian boundary. Paleobiology. 35, 393-412.

Hendy, A.J.W., Kamp, P.J.J. and Vonk A.J., 2009. Late Miocene turnover of molluscan faunas, New Zealand: Taxonomic and ecological reassessment of diversity changes at multiple spatial and temporal scales. Palaeogeography, Palaeoclimatology, Palaeoecology. 280, 275-290.

Hereu, B., Zabala, M. and Sala, E., 2008. Multiple controls of community structure and dynamics in a subtidal marine environment. Ecology. 89, 3423-3425.

Holland, S.M., 2005. The signature of patches and gradients in ecological ordinations. PALAIOS. 20, 573-580.

Holland, S.M. and Patzkowsky, M.E., 2004. Ecosystem structure and stability: Middle Upper Ordovician of Central Kentucky, USA. PALAIOS. 19, 316-331.

Jablonski, D., 1998. Geographic variation in molluscan recovery from the End-Cretaceous Extinction. Science. 279, 1327-1330.

Kowalewski, M., Gurs, K., Nebelsick, J.H., Oschmann, W., Piller, W.E. and Hoffmeister, A.P., 2002. Multivariate hierarchical analyses of Miocene mollusk assemblages of Europe: Paleoceanographic, paleoecological, and biostratigraphic implications. GSA Bulletin. 114, 239–256.

Lafferty, A.G., Miller, A.I. and Brett, C.E., 1994. Comparative spatial variability in faunal composition along two Middle Devonian paleoenvironmental gradients. PALAIOS. 9, 224-236.

Laughlin, D.C. and Abella, S.R., 2007. Abiotic and biotic factors explain independent gradients of plants community composition in ponderosa pine forests. Ecological Modeling. 205, 231-240.

Layou, K.M., 2009. Ecological restructuring after extinction: the late Ordovician (Mohawkian) of the eastern United States. PALAIOS. 24, 118.

Ludwig, J.A. & Reynolds, J.F., 1988. Statistical Ecology. John Wiley and Sons, New York.

McCune, B. and Grace, J.B., 2002. Analysis of Ecological Communities. MjM Software Design, Gleneden Beach, Oregon.

McCune, B. & Mefford, M. J., 1999. Multivariate Analysis of Ecological Data Version 4.25. MjM Software, Gleneden Beach, Oregon.

Minchin, P.R.. 1986. An evaluation of the relative robustness of techniques for ecological ordination. Plant Ecology. 69, 89-107.

Oksanen, J., Blanchet, F.G., Kindt, R., Legendre, P., O'Hara, R.B., Simpson, G.L., Solymos, P., Stevens, M.H.H. and Wagner, H., 2010. vegan: Community Ecology Package. R package version 1.17-3, http://CRAN.R-project.org/package=vegan.

Olszewski, T.D. and Erwin, D.H., 2009. Change and stability in Permian brachiopod communities from western Texas. PALAIOS. 24, 27-40.

Olszewski, T.D. and Patzkowsky, M.E., 2001. Measuring recurrence of marine biotic gradients: a case study from the Pennsylvanian-Permian Midcontinent. PALAIOS. 16, 444–460.

Pandolfi, J.M., 1996. Limited membership in Pleistocene reef coral assemblages from Huon Peninsula, Papua New Guinea: consistency during global change. Paleobiology. 22,152-176.

Patzkowsky, M.E. & Holland, S.E., 1999. Biofacies replacement in a sequence stratigraphic framework: Middle and Upper Ordovician of the Nashville Dome, Tennessee, USA. PALAIOS. 14, 301-323.

R Development Core Team, 2010. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Redman, C.L., Leighton, L.R., Schellenberg, S.A., Gale, C.N., Neilsen, J.L., Dressler, D.L., and Klinger, M.K., 2007. Influence of spatiotemporal scale on the interpretation of paleocommunity structure: lateral variation in the Imperial Formation of California. PALAIOS. 22, 630-641.

Rosenzweig, M.L., 1995. Species Diversity in Space and Time. Cambridge University Press, Cambridge, United Kingdom.

Schindel, D.E., Vermeij, G.J. and Zipser, E., 1982. Frequencies of repaired shell fractures among the Pennsylvanian gastropods of North-Central Texas. Journal of Paleontology. 56, 729-740.

Webber, A.J., 2005. The effects of spatial patchiness on the stratigraphic signal of biotic composition (Type Cincinantian Series; Upper Ordovician). PALAIOS. 20, 37-50.

Zambito IV, J.J., Mitchell, C.E., and Sheets H.D., 2008. A comparison of sampling and statistical techniques fro analyzing bulk-sampled biofacies composition. PALAIOS. 23, 313-321.

Zuschin, M., Harzhauser, M., and Sauermoser, K., 2006. Patchiness of local species richness and its implication for large-scale diversity patterns: an example from the middle Miocene of the Paratethys. Lethaia. 39, 65-78.

**Table 2-1.** List of the six localities sampled with corresponding distance from the northeastern-most locality (Cemetery Road) and latitude and longitude coordinates.

| Locality | Distance from NW | Latitude | Longitude |
|---|---|---|---|
| Cemetery Road | 0 km | 33° 16.1' N | 98° 6.4' W |
| Jacksboro Spillway West | 3.7 km | 33° 14.3' N | 98° 7.3' W |
| Jacksboro Spillway Middle | 3.8 km | 33° 14.3' N | 98° 7.3' W |
| Jacksboro Spillway East | 3.9 km | 33° 14.3' N | 98° 7.3' W |
| Causeway Road | 6.5 km | 33° 13.3' N | 98° 8.8' W |
| Pankey Property | 10.5 km | 33° 10.8' N | 98° 9.1' W |

**Table 2-2.** Analysis of similarity (ANOSIM) results (R-statistics and p-values) for spatial (locality) and temporal (stratigraphic division and relative stratigraphic position) comparisons for the complete dataset and the subset of lower stratigraphic samples.

| | All 29 samples | | Lower 19 samples | |
|---|---|---|---|---|
| | R-statistic | p-value | R-statistic | p-value |
| Locality | 0.06 | 0.208 | 0.29 | 0.009 |
| Stratigraphic Division | 0.56 | < 0.001 | — | — |
| Relative Stratigraphic Position | 0.26 | 0.005 | 0.01 | 0.43 |

**Figure 2-1.** The four sampled localities and specific sample collection spots within and around the area of Jacksboro Texas represented by large circles with stars inside. The small grey circles at the Jacksboro Spillway locality are the three within-outcrop locations sampled at that one locality. Dark grey lines represent roadways, and the light grey area represents water.

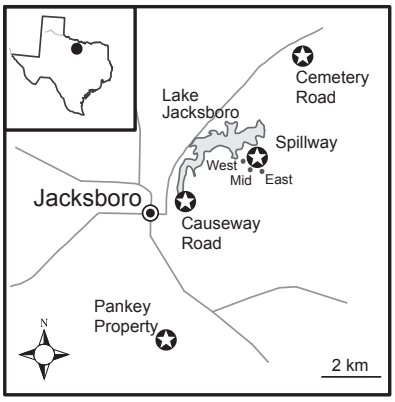


**Figure 2-2.** Non-metric multidimensional scaling (NMDS) ordination of the 29 Finis Shale samples. (A) The sample locality is mapped over the ordination plot. (B) The relative stratigraphic position is mapped over the ordination plot. Stress = 11.5% within two dimensions.

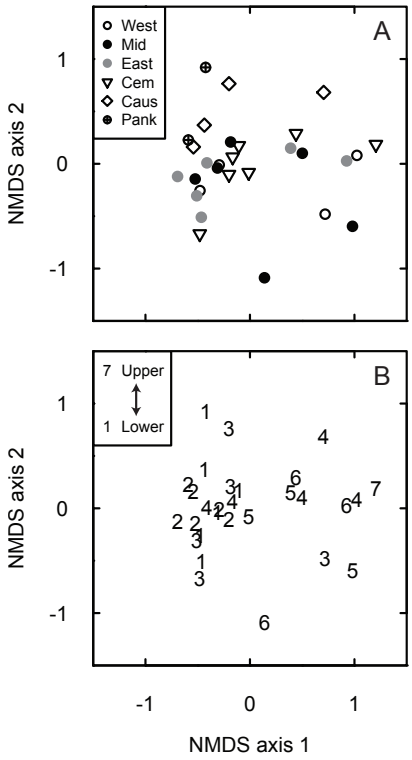


**Figure 2-3.** Non-metric multidimensional scaling (NMDS) ordination of the 19 Finis Shale samples from the lower portions of the stratigraphic sections. The sample locality is mapped over the ordination plot. Stress = 5.4% within two dimensions.

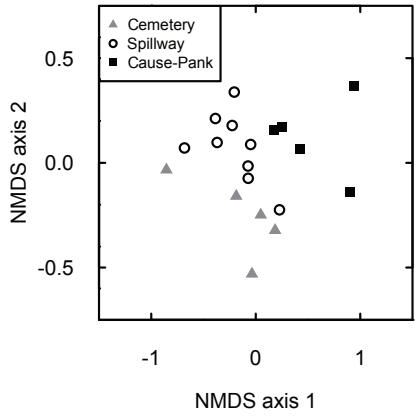
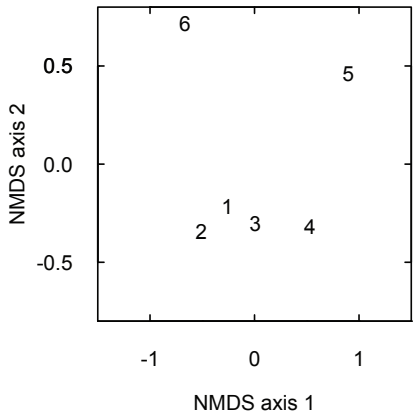


**Figure 2-4.** Non-metric multidimensional scaling (NMDS) ordination of the six samples from the Spillway Middle outcrop of the Finis Shale. Each number (1 to 6) represents the six samples in stratigraphic order. Stress < 1% within two dimensions.



# Chapter 3. The influence of dataset decisions on ecological conclusions: comparison of paleocommunity results between different categorizations of two datasets from Devonian carbonate units[†]

## 1. Introduction

Community paleoecology examines fossil assemblages to study ecological and environmental variation over space and time, yielding insights into the processes that structure ecosystems (Olszewski and Patzkowsky, 2001; Kowalewski et al., 2002; Bonelli et al., 2006; Clapham et al., 2006; Clapham and James, 2008; Heim, 2009). Researchers of community paleoecology have employed an array of methods for collecting, counting, and identifying fossil data. Few studies have quantified the differences that result when various methods are used to examine a single stratigraphic series of communities (see Forcino et al., 2010; Visaggi and Ivany, 2010 for exceptions). Here, I examined fossil assemblages collected from the Rapid Member of the Little Cedar Formation of Coralville, Iowa, USA and from the Waterways Formation of northern Alberta, Canada to compare the paleocommunity results obtained between (1) the use of abundance versus point counts as a counting method and (2) inclusion of all taxa in the paleocommunity versus only the brachiopods. Our goal is to detect any discrepancies between different methods in order to determine when one method may be more appropriate than another in producing accurate representations of the paleocommunities. If two different sets of methods capture different paleocommunity representations, it is paramount to discover the most appropriate methods, which may vary depending upon the type of research. If two different methods produce different paleocommunity results, paleocommunity researchers could base their ecological, environmental, and evolutionary conclusions on flawed results if only using one set of methods; often in past studies, the assumption is made that different methods would not affect the outcome of the study. If a most accurate set of methods can be discovered, relative to a given problem, paleocommunity studies will have increased efficiency, and subsequently, more robust results.

Even if a "best" set of methods cannot be determined (or does not exist), it is important for researchers to know if two sets of methods could produce different paleocommunity results. If applying two sets of methods to the same series of paleocommunities produce different paleocommunity results, researchers may want to explore all possible methods and base any conclusions on the methods that are most suitable to the specific type of question being explored.

Forcino et al. (2010) examined various data categorizations—a subset or restructured dataset to be analyzed based on specific set of parameters of methods—of a paleocommunity

[†] *A version of this chapter is in review for publication. Forcino, F.L., Barclay, K., Schneider, C.L., Linge-Johnsen, S. and Leighton, L.R. (in review) Comparison of the multivariate paleocommunity results from different categorizations of two fossil datasets from carbonate units. Palaeogeography, Palaeoclimatology, Palaeoecology.*

dataset consisting of 13 stratigraphic samples from the late Pennsylvanian Finis Shale of Texas. Here, I use the term "paleocommunity" to refer to a group of fossil assemblages, specifically of marine macroinvertebrates, that in life shared ecological resources. In this sense, a paleocommunity is defined by consistent characteristics (e.g., a similar, common set of constituent taxa) among samples after analysis. The data categorizations compared by Forcino et al. (2010) were calcified biomass (as a proxy for live biomass) versus abundance counts, genus identification versus higher clade identification, and all-taxa versus brachiopod-only analyses. Forcino et al. (2010) obtained the same paleocommunity results (i.e., the same samples consistently grouped together as an identifiable paleocommunity through time) regardless of data categorization method. Two possible conclusions were stated: (1) the use of multivariate methods is sufficiently robust that the choices of collecting, counting, and identifying methods are arbitrary. (2) The results were specific to the Finis Shale, possibly due to lithology, environment, time period, or the dominance of a few brachiopod taxa throughout Finis Shale strata. Here, I continue the efforts of Forcino et al. (2010) by comparing the results derived from the use of various data categorizations on two datasets from a different time (Devonian) and lithology (carbonate) than the Finis Shale. Although Forcino et al. (2010) compared paleocommunity results tallied at the generic level and at a higher clade level, I did not compare paleocommunity analyses at different taxonomic levels in this study as a more complete meta-analysis of this issue was conducted separately (Forcino et al., 2012).

1.1. Abundance versus point counts

A multitude of methods have been used to count fossil material, including point counts (Ausich, 1981; Watkins, 1996; Schneider, 2003), *in situ* abundance counts (Patzkowsky and Holland, 1999; Holland and Patzkowsky, 2004, 2007; Redman et al., 2007), disaggregated abundance counts (Ausich, 1981; Lobza et al., 1994; Olszewski and West 1997), biomass calculations (Staff et al, 1985), biomass measurements (McKinney and Hageman, 2006), calcified biomass (Forcino et al., 2010), and biovolume calculations (Ausich, 1981). The most common methods of counting fossils encased in carbonate units are abundance counts and point counts. In abundance counts, every visible specimen on a surface is counted as one individual, regardless of its size or degree of exposure from the rock. Point counting involves counting only those specimens that are in contact with the intersection points on an overlain grid, and is used as a proxy for biomass as it allows for greater weighting of larger individuals (Watkins, 1988). Larger specimens are likely to touch more points than smaller specimens; therefore, specimens are included in analyses proportionally to their live biomass. However, point counting can result in the omission of taxa, if no individual happens to lie directly under a point. Here, I compared the use of abundance and point counts as a means of quantifying the paleocommunities within the Rapid Member of the Little Cedar Formation and Calumet and Moberly Members of the

Waterways Formation. I then compared the multivariate results (the information and patterns produced based on the taxonomic distributions through and among samples, using multivariate statistical techniques) produced by each counting method.

1.2 All-taxa versus brachiopods-only

Using the Rapid Member dataset, I compared brachiopod-only categorizations to categorizations including all taxa, for two reasons: (1) Many paleocommunity studies examine only one taxonomic group—often the taxonomic group on which the individual researchers are experts—to test hypotheses and determine environmental and ecological conditions of the past. Comparing the paleocommunity signal of brachiopods alone to that of all taxa is one means of evaluating whether ecological and environmental conclusions are accurate when drawn from a limited subset of the assemblage. (2) Because of their abundance and diversity throughout the Paleozoic, brachiopods are often the single taxonomic group used to examine paleocommunity variation (Peters and Bork, 1999; Fall and Olszewski, 2010). If brachiopods alone are sufficient for demonstrating the paleocommunity structure, taxa that are difficult and time-consuming to identify (e.g., crinoids and bryozoans) would not need to be sampled, identified, or counted. If brachiopods alone do not provide sufficient environmental or ecological information, it is important for researchers to know that they need to examine every possible fossil taxonomic group.

2. Geologic Background

2.1 Rapid Member

Samples were collected from the Givetian Rapid Member of the Little Cedar Formation at the Devonian Fossil Gorge in Coralville, Iowa (N 41° 43.3', W 91° 31.9'; Figures 3-1 to 3-2). The Rapid Member was deposited at approximately 20° S latitude during the Givetian as part of transgressive-regressive cycle IIa (Johnson et al., 1985). This transgressive-regressive cycle and others throughout the middle Late Devonian of Iowa can be correlated across North America and Eurasia, suggesting a global eustatic sea level control (Day et al., 2008). The lithology of the Rapid Member at the Devonian Fossil Gorge alternates between argillaceous mudstone to grainstone and massive floatstone (Bunker and Witzke, 1992). Transgression reached its maximum in T-R cycle IIa during the deposition of the lower to middle Rapid Member followed by a regression during upper Rapid Member deposition (Johnson et al., 1985; Day et al., 2008). The lower and middle Rapid Member varies from sparsely to highly fossiliferous, containing a non-reef fauna of brachiopods, crinoids, and bryozoans. During the regression of the upper Rapid Member, corals and stromatoporoids built reefs. These lower and middle intervals range from highly to sparsely fossiliferous. In the upper portions of the Rapid Member that represent relatively low sea level, patch reefs are prevalent, as evident by a biostromal cap at the top of the Rapid Member stratigraphic beds (Bunker and Witzke, 1992).

2.2 Waterways Formation

Samples were collected from the lower Frasnian Calumet and Moberly Members of the Waterways Formation north and east of Fort McMurray, Alberta, Canada from outcrops along the Athabasca and Clearwater rivers (Table 3-1; Figures 3-1 to 3-2). Shale and argillaceous limestone of the Waterways Formation accumulated as prograding clinoforms that filled the eastern portions of the Alberta Basin. Along the shallow, northeastern margins of the basin, the clinoform tops formed an extensive carbonate platform that was interrupted by periods of increased shale deposition. The Calumet Member, which is in the lower Waterways Formation, is an argillaceous and fossiliferous limestone (Norris, 1963). The Moberly Member in the upper Waterways Formation contains predominantly fossiliferous argillaceous to non-argillaceous limestone (Norris, 1963; Buschkuehle, 2003) and is heavily bioturbated in some units.

A primary difference between the Rapid Member and the Waterways Formation is the relative rarity of bryozoans and corals in the communities collected in the Waterways Formation in the present study. Samples used herein were from the level-bottom communities common in the Calumet and lower Moberly, rather than from the biostromes present in the upper Moberly. This difference in taxonomic composition allows an additional test of the effect that abundant colonial and easily disarticulated taxa may have on the resulting correlations between abundance and point count categorizations in samples of similar lithology and age.

## 3. Methods

### 3.1 Rapid Member sample collection

The Devonian Fossil Gorge in Coralville Iowa consists of a 300 meter-long outcrop of terraced limestone beds. Samples were collected in-place from the 15 least weathered horizons, spanning 8.7 m from the contact between the Rapid Member and the underlying Solon Member. On each surface, fossils were counted using two methods: (1) abundance counts were tallied to a minimum of 200 specimens per sample and (2) point counts were conducted using a 605 $cm^2$ grid with lines 1 cm apart. At each intersection of the grid, a count of one was tallied for the particular taxon occupying the intersection point; if two different taxa were under the same intersection point (i.e., one specimen is on top of the other), each taxon received a count of one. If there was no specimen under the intersection, nothing was added to the count. In addition, all specimens within the total grid space were counted to get both abundance and point counts within a constant surface area. The abundance data categorization taxon-sample matrix consisted of 3124 specimens from 39 genera. The point count data categorization taxon-sample matrix consisted of 803 point counts from 39 genera.

### 3.2 Waterways Formation sample collection

Samples were collected from 13 outcrops located along the Athabasca and Clearwater Rivers near Fort McMurray, Alberta, Canada (Table 3-1; Figures 3-1 to 3-2). At each outcrop, one to eight limestone slabs were collected for fossil identification and counting in the laboratory. A total of 22 limestone slab samples were collected. All fossils specimens were

identified to the genus level and all visible specimens were counted for the abundance dataset. This taxon-sample matrix, based on the abundance categorization, consisted of 2365 specimens from 20 genera.

Only 11 of the 22 limestone slabs were point counted. The other 11 samples could not be point counted because the fossils were too sparse on the limestone surface (< 20 per slab). The typical area point counted was 605 $cm^2$. However, if the fossils were abundant enough for analysis, but were sparsely distributed across the surface, the grid area was increased to capture a sufficient abundance of fossils. The point count data categorization taxon-sample matrix consisted of 670 point counts of 13 genera.

For both the Rapid Member and the Waterways Formation abundance count datasets, colonial taxa (bryozoans and tabulate corals) and easily disarticulated taxa (crinoids) were included as one individual for every 1 cm of length, using the maximum length of each fossil fragment. This technique has been used in previous studies (Patzkowsky and Holland, 1999; Forcino and Leighton, 2010) as a means of equating the abundances of colonial and easily disarticulated taxa to those of solitary taxa (brachiopods and mollusks). Since I did not find any crinoid calyx material or any other means to distinguish different genera of any echinoderm group, I classified all echinoderm fossils based on morphology and distinguishing characteristics (e.g., columnal shape, size and shape of the lumen). Although these features may not always indicate a distinct genus, I made the assumption that they were different enough to represent different portions of a paleocommunity.

3.3 Analytical methods

I divided the Rapid Member dataset into four data categorizations: (1) abundance of all-taxa, (2) point counts of all-taxa, (3) abundance of brachiopod taxa, and (4) point counts of brachiopod taxa. The Waterways Formation dataset produced two data categorizations: (1) abundance of all-taxa and (2) point counts of all-taxa (there were no brachiopod-only versus all-taxa categorizations because the dataset was comprised almost exclusively of brachiopods).

Paleoecological studies often transform or standardize community data in order to account for differences in sample size or collection methods. Here, I conducted statistical analyses on non-transformed, absolute abundance data and on percent-transformed community data. Percent transformation is a means of accounting for differences and biases in collection methods or sample size. It is often done when studies consist of both bulk-collected shale samples and surface-counted carbonate samples. The results for the analyses conducted on percent transformed datasets are included in the results section for completeness. However, I focused on the non-standardized results for our interpretation for two reasons: (1) I followed a consistent sampling protocol with sample sizes that were relatively similar, especially the Rapid Member dataset (from 203 to 219; and 18 to 390 for the Waterways Formation dataset). Point count surface areas were held constant for the Rapid Member dataset, and the differences in

surface area of the point count samples within the Waterways Formation dataset did not cause any noticeable differences between standardized and non-standardized ordination results. (2) Since the purpose of this paper was to compare differences that may arise from different methodological choices, I did not want to base our interpretations on results from methods that may smooth out differences, as standardizing may.

For both datasets, taxa or samples with low numbers (total sample abundance or rare taxa with < 5 total specimens) were retained because I did not want to alter the data prior to comparisons. If rare taxa and low-abundance samples were removed, the abundance and point count categorizations would be more similar and possibly artificially increase the correlation statistics between the two categorizations.

3.4 Statistical methods

The comparisons of data categorizations were conducted using three multivariate statistical methods:

(1) Using the vegan package in R 2.14 (Oksanen et al., 2011; R Development Core Team, 2011), Mantel Tests of correlation were performed between the Bray-Curtis dissimilarity matrices (measures of the differences between each object in a taxon-sample matrix) for each of the data categorization comparisons. The Mantel Test tests the similarity of two matrices of dissimilarity indices by permuting each of the elements in the dissimilarity matrix 999 times, to derive a distribution of correlation values (Mantel, 1967; Fall and Olszewski, 2010). The resulting R-statistic is analogous to the Pearson's Product Moment Correlation Coefficient (r); with increasingly similar data matrices, the Mantel R-statistic will approach 1.

(2) For each of the data categorization datasets, non-metric multidimensional scaling (NMDS) ordinations of the samples were performed using the Bray-Curtis dissimilarity index (Clarke and Ainsworth, 1993; Bush and Brame, 2010; Figures 3-3 to 3-4). Specifically, the "metaMDS()" function in R was used with two dimensions and autotransformation off. For each of the categorization comparisons, the Pearson's Product Moment Correlation was conducted between NMDS axis-one scores.

All ordinations were run examining the taxonomic distributions among samples. Ordination is an exploratory multivariate visualization tool that allows multidimensional relationships of samples to be examined in a low number of dimensions (Legendre and Legendre, 1998; McCune and Grace, 2002). Because ecological datasets contain samples with taxonomic objects, each with some abundance, ordination is the standard way to visualize the similarities and differences among samples or taxa. Samples that have more similar taxonomic distributions plot closer together in the ordination space.

NMDS ordination iteratively searches for a best-fit solution between the rank dissimilarity indices and the distribution of samples in a low dimension ordination space. This non-parametric approach is appropriate for community data, which are typically non-normally

and non-linearly distributed (Bush and Brame, 2010). The best-fit solution is assessed by the stress of the ordination; low stress represents a better NMDS solution (Kruskal, 1964). All ordination methods have advantages and disadvantages. NMDS was used here because it is widely accepted among ecologists and paleoecologists, it uses a fixed number of dimensions, and it is non-parametric (Bush and Brame, 2010).

(3) Procrustean Randomization Tests (PROTEST) were performed comparing procrustes transformed ordinations of higher- and lower-taxonomic levels (Jackson, 1995; Peres-Neto and Jackson, 2001). NMDS does not always assign the maximum explanation of variation in the ordination space to the first axis. Moreover, two different ordinations might not appear to be similar at first because they are close reflections to each other in ordination space. To address these possibilities, the first step in PROTEST is to perform a Procrustes transformation, which minimizes the sum-of-squares deviations between the two ordination results through translation, reflection, rotation, and dilation. Thus, the two ordination results are reoriented such that they are aligned as closely as possible in ordination space, which permits a more accurate assessment of similarity. The residuals between the two ordinations post-transformation are calculated and produce the $m^2$-value. The $m^2$-value is similar to the r-value resulting from a Pearson's Product Moment Correlation; the closer $m^2$ is to 1, the more similar the two ordinations. Subsequent to the Procrustes transformation, PROTEST randomly permutes the ordination scores for all samples for 999 iterations, and the $m^2$-value is calculated for each iteration; a realized p-value, indicating the significance of the $m^2$-value, is then calculated by determining the percentage of iterations in which the $m^2$-values from the randomized iterations is greater than the $m^2$-value for the actual dataset.

3.5 Visual assessment

I also visually examined each of the ordination results for all of the categorization comparisons to assess the similarity of the multivariate results of all the dataset comparisons in a completely different manner from the statistical tests. Although this comparison method is less quantitative, and potentially subjective, visual examination of each individual ordination comparison provides an independent approach that may potentially capture details that would not be recognizable by the Mantel Test, NMDS axis-one, or PROTEST goodness-of-fit tests alone. Although a $p < 0.05$ may indicate a significant level of overall similarity between two paleocommunity results when using these goodness-of-fit tests, the resulting goodness-of-fit value does not necessarily indicate whether two results would be interpreted the same way by a researcher. Most community paleoecologists are not interested so much in the exact position of a sample, or the distance between two samples, in ordination space. In many cases, the interpretation of an ordination plot hinges on the identification of discreet clusters of samples in ordination space or of a gradient of samples. While two ordinations may be statistically similar, these sorts of details (e.g., membership within a specific cluster of points) are not explicitly

evaluated by any of the more sophisticated statistical methods. Thus, it is possible for a test to produce a significant result, but a few key samples may change position radically, which would be identified when examined visually. Similarly, an insignificant result might still reveal a critical pattern. Changes in the position of a few samples in ordination space may be sufficient to result in a different interpretation of ecological controls on sample distribution. The situation is analogous to visually inspecting a scatterplot when performing correlation statistics. Most workers have observed that a relatively weak correlation may still be associated with a significant p-value, partly because of the size of the dataset, thus emphasizing the need to examine the associated graphic result.

Although each categorization comparison was visually assessed based on the unique differences that arose in each set of ordinations, I attempted to adhere to a set of standards when deciding if two ordination results are the "same" or "different". I interpreted two ordination results to be the same if fewer than ~10 % of the sampling units show change in position (less than 50 % the length of one or both axes). When two ordinations are classified as the same, I infer that the groupings and patterns of the sampling units within the ordinations do not change enough to lead to different paleocommunity interpretations and conclusions. I interpreted two ordinations results as different if greater than 10 % of the sampling units show change in position. This infers that there was sufficient movement in samples within ordination space to potentially lead to different paleocommunity interpretations and conclusions. This assessment was conservative; if I were unsure whether researchers would interpret the ordination results as the same, I labeled the comparisons as not the same.

The above visual assessment technique does not take into account possible independent data that a researcher may map onto the ordination. Ordination itself is not a statistical test. Researchers examine the distribution of samples in ordination space and compare the independent variables (e.g., geochemical or sedimentological data) of each of the ordination samples by mapping these variables onto the ordinations. For example, if a researcher collects 10 fossil community samples from two depths, the researcher may want to ordinate the 10 samples and map on the two depths. If there is strong separation of the samples in ordination space based on depth, this is evidence that depth is controlling the taxonomic composition and abundance of the samples.

In order to examine the effect that different categorizations may have on the interpretation of mapped independent variables, I created a series of each of the different ordinations of the categorization comparisons with various selected groupings of points (Figures 3-5 to 3-10). The different groups of points selected in each of the comparisons are meant to represent various independent data that may be mapped onto the ordinations. For each of the six data categorization comparisons (four for the Rapid Member dataset and two for the Waterways Formation dataset), I selected three different hypothetical independent variable schemes. These

groupings were selected based on locations where there was apparent separation of points. For example, if there was a group of four points that grouped in the upper half of the ordination, these four points would be selected as one group and the other points would be a different group. I mapped the same two variables on each of the two ordinations that were being compared and then examined whether the same clusters, groupings, or patterns exist within both of the categorizations (Figures 3-5 to 3-10). If the same clusters, groupings, or patterns exist in all three of the mapped independent data visual examinations, this would be evidence that the categorizations are producing the same paleocommunity results; the same interpretations would be made regardless of the independent variable mapped. Conversely, if the clusters, groupings, or patterns differ between categorization comparisons, this would be evidence that the categorizations are producing different paleocommunity results; different interpretations would result between the different categorizations based on the independent variables.

I did not compare the results of the various categorizations with measured independent environmental or ecological data (e.g., lithology or geochemical data) mapped onto the ordinations. The ordination comparisons are meant to examine the results of the categorizations without regard to how the results specific to the Waterways Formation or Rapid Member would be interpreted. This study is meant to help others examine data in other studies without regards to the particulars or questions under investigation in any one study. So, I wanted to take an approach that did not required ecological or environmental interpretation of our fossil data. Those data will be presented in subsequent publications. In addition, any interpretation I may have based on independent variables within the Waterways Formation or Rapid Member may differ from other researchers.

## 4. Results

### 4.1 Abundance versus point counts

Among the Rapid Member comparisons of abundance and point count categorizations (raw data and percent transformed data, complete abundance area versus point count area, and within the point count area comparisons), the three different multivariate correlation methods (i.e., Mantel Test, NMDS axis-one correlation, and PROTEST) produced a range of goodness-of-fit statistics (i.e., Mantel Test R-statistics, NMDS axis-one correlation r-values, and PROTEST $m^2$-values) from 0.61 to 0.78 ($p < 0.01$; Table 3-2). Because results between the transformed and non-transformed datasets were the same, and to limit confusion (e.g., the number of different types of categorizations discussed that have similar sounding labels), the results and discussion will only address goodness-of-fit statistics for non-standardized categorizations. I visually assessed the two ordinations as different due to changes in position of most sample points (Figure 3-3A). In addition, points that were close to one another do not stay constant between the two ordinations; different pairings and clusters exist between the two ordinations. Specifically, within the exercise of examining the effect of mapped independent data, the three different

examples resulted in three slightly varying levels of difference between the categorizations (Figure 3-5). The first grouping example had similar separation between the two variables within both categorization comparisons (Figures 3-5A and 3-5D). The second two grouping led to some ambiguity in the separation of the mapped variables (Figures 3-5B, 3-5C, 3-5E, and 3-5F).

Among the Rapid Member comparisons of abundance and point count categorizations for brachiopods-only, goodness-of-fit statistics ranged from 0.63 to 0.84 ($p < 0.01$; Table 3-2). The brachiopod-only ordinations of the Rapid Member categorizations were more similar to one another than the ordinations of the categorizations that included all taxa (Figure 3-3C). However, I visually assessed the two ordinations as different. All points, except for two, change position along axis two, and two points change along axis one. This contrary motion is the reason these two ordinations were assessed as different. Within the examination of mapped independent data, one of the three sets of comparisons had the same cluster separations (Figure 3-6A and 3-6D). However, the remaining two comparisons resulted in differences in the separation and clustering of the two mapped variables (Figure 3-6).

Among the Waterways Formation dataset, comparisons of abundance and point count categorizations resulted in goodness-of-fit statistics ranging from 0.73 to 0.90 ($p < 0.001$; Table 3-3). I visually assessed the two ordinations as different because all but one point changes position (Figure 3-4A). Seven points change along axis two, and two points change along axis one. Within the examination of mapped independent data, two of the three example clusters remain fairly separated in both categorizations (Figure 3-7). However, one of the examples leads to changes in the clustering (Figures 3-7C and 3-7F).

Among the Waterways Formation comparisons of abundance and point count categorizations without crinoids, goodness-of-fit statistics range from 0.85 to 0.93 ($p < 0.001$; Table 3-3). Although there is less difference between these two ordinations than the Waterways dataset abundance versus point counts with crinoids, I visually assessed the two ordinations as different (Figure 3-4B). All sample points change position. Within the examination of mapped independent data, all three examples resulted in some difference in the clustering separation of the two mapped variables (Figure 3-8).

4.2 All-taxa versus brachiopods-only

The goodness-of-fit statistics for the non-standardized Rapid Member comparisons of all-taxa versus brachiopods-only, when abundance was used to measure fossil material, ranged from 0.59 to 0.88 ($p = 0.02$ to $p < 0.001$; Table 3-3). I visually assessed the two ordinations as different because the samples displayed variation between corresponding sample points along axis-one, axis-two, and diagonally, with no consistent direction of change (Figure 3-3B). Within the examination of mapped independent data, the clusters remain similar within all of the comparisons (Figure 3-9). However, one point does change clusters in each of the three examples.

The correlation statistics for the Rapid Member comparisons of all-taxa versus brachiopods-only, when abundance was used to measure fossil material, ranged from 0.30 to 0.64 ($p = 0.18$ to $p < 0.001$; Table 3-3). This comparison led to the most variation between two ordinations. I visually assessed the two ordinations as different because all points change along axis-one or axis-two, and many change diagonally, with no consistent direction of change (Figure 3-3D). Within the examination of mapped independent data, all three examples resulted in different clusters and patterns of the mapped independent variables (Figure 3-10).

## 5. Discussion

### 5.1 Interpreting multivariate results

The goodness-of-fit statistics for the comparisons of abundance versus point counts ranged from 0.61 to 0.88 ($p < 0.01$; Tables 3-2 and 3-3). Although the p-values were less than 0.001 and 0.01 for all of the Mantel Tests and PROTESTS, respectively, these p-values may not directly address the question of interest (would these ordinations be interpreted in the same way?) in the same manner as a p-value for a bivariate Pearson's Product Moment Correlation. While a $p < 0.05$ combined with a moderate (0.5 to 0.8) PROTEST $m^2$-value or Mantel R-statistic indicates significant overall similarity of data categorizations (in the sense that there is a 95% confidence that the result is not obtained by chance), it does not necessarily indicate similarity of details that would lead workers to a consistent interpretation of two ordinations. In this regard, even a moderate goodness of fit statistic may be too low for two ordinations to be interpreted as the same. The p-values in both of these tests are calculated by determining if the goodness-of-fit values differ significantly from randomized permutated values. So, both the goodness-of-fit statistics and the p-values for both the Mantel Test and the PROTEST do not have a definitive value that can be considered a threshold between two datasets being interpreted as the same or different.

Because of the lack of a known threshold value, these tests provide important information, but they may not provide information about how researchers would interpret the two results being compared. Further information is needed fully to understand the similarity between two categorizations. I conducted a complementary visual assessment to further evaluate the various categorization comparisons. I interpreted all of the categorization comparisons to be different; the patterns and groupings within the raw ordination results differ (Figures 3-3 and 3-4).

In some cases, the two categorizations produce the same separation of the samples representative of two different mapped variables (Figures 3-5 to 3-10). However, since there are one or two examples that result in different clusters and patterns for each of the six categorization comparisons, I interpret these changes in the clustering and patterns of the mapped independent variables as differences arising between the different data categorizations. This

further supports our claim that the different categorizations examined produce paleocommunity results that would be interpreted differently.

The ultimate purpose of multivariate statistical techniques such as ordination is to use the ordination results to test independent variables (e.g., water depth, predation intensity) that may be controlling community distribution or composition. Based on the differences found between the pairs of categorizations, the interpretation of the primary controlling variable could vary greatly between the use of the two methods. The patterns and groupings of samples change drastically enough that the same independent variable would be unlikely to correlate with the multivariate results from all categorizations.

I recognize that our visual assessment, and even our choice of criteria for identifying "same" or "different", is subjective and that other workers might potentially interpret pairs of ordinations differently. However, the very fact that there are multiple cases in which I might interpret two ordinations very differently, despite a low p-value indicating great significance, also must raise some questions about how researchers use and interpret these multivariate tests; the results suggest that low p-values for the multivariate tests do not inherently imply consistency of interpretation. Our visual examination suggests that goodness of fit statistics may need to be very high (i.e., > 0.85) to ensure that two ordinations would be interpreted in the same way – even if they display general similarity (Figure 3-3C). Further research on this problem needs to be done (Forcino et al., in review; Forcino and Leighton, 2012), but for the moment, caution must be used when interpreting the quantitative multivariate results.

5.2 Abundance versus point counts of all taxa

The paleocommunity patterns resulting from using abundance as a means of counting fossils produces consistently different paleocommunity patterns than those resulting from the use of point counts. This was consistent between both the Waterways Formation and the Rapid Member datasets using multivariate goodness-of-fit tests as well as a visual assessment. The goodness-of-fit statistics for the comparisons of abundance versus point counts ranged from 0.61 to 0.88 ($p < 0.01$; Tables 3-2 and 3-3). The PROTEST $m^2$-values ranged from 0.62 to 0.88, and the Mantel R-statistics ranged from 0.59 to 0.85.

Forcino et al. (2010), examining the Finis Shale, found a strong correlation of NMDS axis-one scores between ordination results of abundance counts and ordination results of calcified biomass. Forcino et al. (2010) did not employ the other methods used in the present study, so the only direct comparison of results is for correlation of NMDS axis-one scores (Table 3-4). In contrast, within the two carbonate units, abundance and point count categorizations produced different results. Within all of the correlations of NMDS axis-one scores for the three different units (Finis Shale, Waterways Formation, Rapid Member), the Finis Shale had greater r-values than the two carbonate units (Table 3-4).

One possible cause for the difference between counting methods observed in the present study could be that the present study used point counts as a biomass proxy whereas Forcino et al. (2010) examined calcified biomass as a proxy for biomass, an approach that is not feasible for specimens in limestones. Fossils encased in limestone are difficult to impossible to extract without destruction of details required for identification. Furthermore, it is not possible to conduct point counts in siliciclastic shale units such as the Finis Shale. Any one surface is not nearly as fossiliferous as the limestone surfaces examined in the Waterways Formation and Rapid Member. Fossils within the shale must be disaggregated and tallied. The process of weighing all of the fossils in the Finis Shale to determine calcified biomass may have more completely represented the communities, similar to the abundance counts, because all of the taxa are included completely; the complete mass of every specimen was included. Conversely, point counts may be less similar to abundance because each specimen is only included based upon the number of points it happens to intersect. There is always a chance that larger taxa may only lay within one intersection. In this manner, point counts are not as all inclusive as abundance and biomass. So, this difference in the selection of the biomass-proxy could have led to stronger correlations in the Forcino et al. (2010) study compared to the present study.

The difference in lithology between the present study and Forcino et al. (2010) is another possible cause for the differences that arose between counting methods. Both the Waterways Formation and the Rapid Member are carbonate units. Forcino et al. (2010) found strong similarity among counting methods within the siliciclastic Finis Shale. Environmental or taphonomic factors that differ between carbonate and siliciclastic units may have ultimately led to the differences found between these two studies. For example, there may be differences in the taxa that are able to survive in a more turbid environment (such as those in which shale would be deposited) compared with an environment with lower clastic sediment input, such as those that form a limestone. Corals and many suspension feeders are not able to filter out large amounts of sediment and may not be found in environments that are represented by shale in the rock record. Although previous research demonstrated low similarity between community results using abundance and biomass within siliciclastic units (Morris, 1985; Bush et al., 2007), these studies do not definitely refute the possibility that lithology has some control on the required methods for paleocommunity research. Since these studies do not examine any carbonate units, they do not provide information regarding factors that may consistently lead to differences among counting methods.

The stratigraphic section of the Finis Shale examined by Forcino et al. (2010) was homogeneous, particularly compared to the Waterways Formation and Rapid Member stratigraphic sections examined for the present study. Within the Finis Shale, it is not visually obvious that any of the stratigraphic horizons from which fossils were sampled are any different from any other, except for some minor variation in silt content. The entire stratigraphic extent of

the Finis Shale is dark grey to black, non-fissile shale. However, the Rapid Member varies in lithology from argillaceous, recessive limestone to more massive, resistant units to some thin fissile units. Similarly, the Waterways has some very thin, argillaceous, recessive beds as well as some massive beds, along with a few thin shale units. Moreover, there was greater community homogeneity within the Finis Shale as well. There are only two different communities within the entire Finis Shale (Forcino et al., 2010; Forcino et al., 2012). Here, there are at least three different communities within the Waterways Formation, and perhaps even more within the Rapid Member (Figures 3-3 and 3-4; Schneider et al., 2012). Whether this homogeneity was caused by environmental stability or some other factor, there is less variation throughout the Finis Shale depositional environment than in the Rapid Member and the Waterways Formation carbonate settings. This heterogeneity within the Waterways Formation and Rapid Member may have led to the differences between the abundance and point count categorizations.

If the taxonomic or lithological (environmental) factors do not account for the differences between abundance and point count results, it is possible that biological factors caused these differences. If this is the case, which counting method is the more accurate means of assessing the paleocommunity signal? Many workers view proxies of biomass or biovolume such as point counts as more biologically meaningful indicators of community structure (Watkins, 1996). Abundance counts exhaustively include all whole or nearly whole fossil specimens. However, there is no definitive means of equating colonial and easily disarticulated taxa with solitary taxa. Although several methods have been employed (e.g., equating 2 cm length to 1 individual), no single technique is universally accepted to cover questions of taxonomic comparisons. If a method of equating colonial and easily disarticulated taxa to solitary taxa is found and deemed appropriate for the research question under investigation, this may lead to more similar results between abundance and point counts. However, until that method is determined, it may be best to conduct both abundance and point counts. When under resource constraints, point counts are recommended. When both methods (abundance and point counts) yield different results, I recommend that the point count results be given more weight, based on the more equitable inclusion of solitary, colonial, and easily disarticulated taxa. However, abundance counts are recommended if the communities in carbonate units are to be compared with communities for which point counts cannot be obtained (e.g., communities in highly friable sandstones or shale, where bedding surfaces are most often destroyed before examination).

If different results are obtained when using two or more methods, there may be an inherent problem with the data. Although point counting may provide a more biologically meaningful result, if the results obtained by point counting are inconsistent with results obtained by a different method (e.g., abundance counts), this may mean that the samples are not representative of either the complete fossil community or the once living community. In this manner, abundance counts are important to examine in addition to point counts because if the

analyses based on abundance counts and the analyses based on point counts are consistent with one another, the paleocommunity results are more robust. Thus, a researcher can be more certain of their paleocommunity interpretations and ecological conclusions.

5.2.1 Colonial and easily disarticulated taxa effect

Crinoids made up a greater percentage in the point count categorization compared to the abundance categorization (Figure 3-5). This possibly led to differences between the ordinations, similar to that seen in Watkins (1996), where differing paleocommunity results were attributed to the high percentage of crinoids and bryozoans in the dataset. As described previously, to test the negative effect of colonial and easily disarticulating taxa on the goodness-of-fit between the abundance and point count categorizations, I culled bryozoans and crinoids (same categorization as brachiopods-only for the Rapid Member dataset), and re-ran multivariate comparison analyses (Figures 3-3D and 3-4C). If the bryozoan-crinoid component of the abundance and point count datasets caused the differences between the paleocommunity results, the goodness-of-fit statistics should increase when these groups are culled. As expected, omission of bryozoans and crinoids increased the mean goodness-of-fit statistic (the average of all goodness-of-fit statistics from all the different comparison methods) from 0.61 to 0.78 (Tables 3-2 and 3-3; Figures 3-3 and 3-4). Thus, when crinoids and bryozoans were culled from the analyses of the Rapid Member and Waterways Formation datasets, the abundance and point count paleocommunity results became more similar.

The fact that our abundance and point count multivariate results were more similar when colonial and easily disarticulated taxa were culled demonstrates the importance of including colonial and easy disarticulated taxa in analyses along with solitary taxa. When taxa such as crinoids and bryozoans are included in the data, point counts capture a paleocommunity that differs markedly from that based on abundance counts because the colonial and easily disarticulated taxa have an effect on how the paleocommunity is perceived. Furthermore, although there were limited differences among categorizations of the Finis Shale examined by Forcino et al. (2010), a minor difference that arose was between abundance and biomass categorizations of higher stratigraphic samples. These samples contained a greater amount of crinoids and bryozoans. Since the biomass measurement allowed for a means of including all taxa equally, the presences of the bryozoans and crinoids in the higher stratigraphic samples most likely caused the differences between the categorizations of the samples. Therefore, depending upon the question the researcher is exploring and the availability of fossil material, for the most accurate representation of a community, researchers should use methods that include colonial and easily disarticulated taxa. Even if researchers are interested in examining ecological variation of a single taxon or one taxonomic group, it is important also to examine the complete picture of what is occurring in the system. The researchers and the paleontological community would benefit from an understanding of how taxonomic groups align with the ecological patterns

revealed by all fossil taxa. Studies about how brachiopod communities change through an environment perturbance are interesting, but if these data were different from the overall understanding and patterns of analyses including all organisms through that same perturbance, this would be useful information for examining alternate hypotheses and interpretations. All paleontological endeavors provide useful and interesting information. However, for these reasons, I recommend a consistent, all-encompassing approach that includes all fossil taxa.

Variation in biovolume among different taxa may cause differences in the paleocommunity signals when conducting point counts (specifically for the Rapid Member dataset). In contrast with abundance counts, point-counts allow larger taxa, which might be expected to play a greater role in acquisition of resources within the community, to carry greater weight in the the analysis than smaller taxa. Determining a taxonomic size effect is outside the scope of the present study, but it will be a part of future work for the authors.

5.3 All-taxa versus brachiopods-only

The correlations between the use of all-taxa versus brachiopods-only for the abundance counts of the Rapid Member dataset ranged from 0.59 to 0.88 ($p < 0.02$). When using point counts, there was a marked difference between the brachiopod-only and all-taxa categorizations, with correlation statistics ranging from 0.30 to 0.64 ($p < 0.18$; Table 3-2; Figure 3-3). I interpret these weaker correlations, as well as the visual differences, to indicate probable different paleocommunity results. Therefore, taxa other than brachiopods can drastically influence the community structure of a dataset. Many factors may have influenced the poor goodness-of-fit between the Rapid Member brachiopod-only and all-taxa categorizations. The Rapid Member does contain corals and stromatoporoids. *Hexagonaria*, a colonial coral, achieves great abundance higher in the Rapid Member. This may indicate a narrow range of environments represented by the Rapid Member. Non-brachiopod taxa may have a greater influence on the community signal because these taxa have the most influence on the apparent differences of the communities in ordination space.

Although brachiopods are often easier to identify, other taxa should be collected and included in analyses in order to ensure a more complete community characterization and interpretation. One could argue that brachiopods alone can be collected when the study is solely focused on the brachiopod assemblage; the brachiopods alone may track some ecological or environmental variable that is obscured when the community is examined as a whole. However, brachiopods usually represent only a portion of a broader paleocommunity. The information provided by other fossil taxa within a community may be necessary for understanding variation in the brachiopod-only component of the community. Moreover, the brachiopod-only ordination pattern is not necessarily a subset of the pattern from the full community; it could be completely different. I recommend using the most complete fossil community as possible when conducting

paleocommunity research, recognizing that logistic difficulties may sometimes make such research impossible.

## 6. Conclusions

Although Forcino et al. (2010) found a strong correlation in the paleocommunity results between those obtained using abundance counts and those using calcified biomass, the comparison of abundance and point counts within the Rapid Member and the Waterways Formations, and from numerous other studies (Ausich, 1981; Staff et al., 1985; Morris, 1986, Watkins, 1996; Bush et al., 2007), demonstrate strong differences between paleocommunity results obtained using abundance and those using point counts (or other proxies for biomass). Using NMDS axis-one score correlations, the Finis Shale had consistently greater r-values than the Waterways Formation and Rapid Member (Table 3-4). As evident from the increased correlation statistics of the paleocommunity results between the abundance and point counts of the Waterways Formation dataset when crinoids were culled, colonial and easily disarticulated taxa can have a strong effect on the ordination results obtained from point counts. In communities that consist of mostly solitary taxa, like brachiopods, abundance counts may be sufficient for counting fossil material. However, because abundance and point counts result in different ordination results, conducting point counts in concert with abundance counts is recommended when resources permit. If the same result is found with both methods, it verifies that the interpretations derived from the paleocommunity analysis are valid. If the results differ, as long as the point count grid sufficiently includes taxa of all sizes, I recommend using the point count result, as it takes into account the calcified biovolume of fossilized taxa and allows equal inclusion of non-solitary taxa.

Brachiopods alone do not produce the same paleocommunity result as when all taxa are included in the analyses. There is a poor correlation between the paleocommunity results of the Rapid Member dataset categorizations using all-taxa and using brachiopod taxa only – brachiopods are not automatically good indicator taxa. Restricting analysis to a single clade, largely for logistic rather than biological reasons, is not recommended. In order to get a full and accurate view of the paleocommunity under study, I recommend the identification and quantification of all taxa for environmental, ecological, and evolutionary interpretations drawn from paleocommunities.

## 7. Acknowledgements

This research would not be possible without funding to Forcino from the University of Alberta Circumpolar Research Institute and the Geological Society of America Student Grant, and funding to Leighton and Schneider from NSF- EAR-0746072. Thank you to Emily Stafford and Brian Chatterton for their input to this research, and the Alberta Geological Survey for additional resources and input. The manuscript was improved greatly by a review from Mark Patzkowsky.

8. Bibliography

Ausich, W.I., 1981. Biovolume revisited: a relative density Index for paleoecological analyses. Ohio Journal of Science. 81, 268–274.

Bonelli Jr., J.R., Brett, C.E., Miller, A.I. and Bennington, J.B., 2006. Testing for faunal stability across a regional biotic transition: quantifying stasis and variation among recurring coral-rich biofacies in the Middle Devonian Appalachian Basin. Paleobiology. 32, 20–37.

Bunker, B.J. and Witzke, B.J., 1992. An upper Middle through Upper Devonian lithostratigraphic and conodont biostratigraphic framework of the Midcontinent Carbonate Shelf area, Iowa, in: Day, J., Bunker, B.J. (Eds.), The Stratigraphy, Paleontology, Depositional and Diagenetic History of the Middle-Upper Devonian Cedar Valley Group of Central and Eastern Iowa. Iowa City, Iowa, Iowa Department of Natural Resources, 3–26.

Buschkuehle, B.E., 2003. Sedimentology and stratigraphy of Middle and Upper Devonian carbonates in northern Alberta: A contribution to the carbonate-hosted Pb-Zn (MVT) targeted geoscience initiative. Alberta Energy and Utilities Board, EUB/AGS Geo-Note, 2002–2014.

Bush, A.M., Kowalewski, A., Hoffmeister, A.P., Bambach, R.K. and Daley, G.W., 2007. Potential paleoecological biases from size-filtering of fossils: strategies for sieving. PALAIOS. 22, 612–622.

Bush, A.M. and Brame, R.I., 2010. Multiple paleoecological controls on the composition of marine fossil assemblages from the Frasnian (Late Devonian) of Virginia, with a comparison of ordination methods. Paleobiology. 36, 573-591.

Cadotte, M.W., Cardinale, B.J. and Oaklet, T.H., 2008. Evolutionary history and the effect of biodiversity on plant productivity. Proceedings of the National Academy of Sciences. 105, 17012–17017.

Clapham, M.E., Bottjer, D.J., Powers, C.M., Bonuso, N., Fraiser, M.L, Marenco, P.J., Dornbos, S.Q. and Pruss, S.B., 2006. Assessing the ecological dominance of Phanerozoic marine invertebrates. PALAIOS. 21, 431–441.

Clapham, M.E. and James, N.P., 2008. Paleoecology of Early-Middle Permian marine communities in eastern Australia. response to global climate change in the aftermath of the late Paleozoic ice age. PALAIOS. 23, 738–750.

Clarke, K.R. and Ainsworth, M., 1993. A method and linking multivariate community structure to environmental variables. Marine Ecology. 92, 205–219.

Day, J., Witzke, B.J. and Bunker, B.J., 2008. Overview of Middle and Upper Devonian Cedar Valley Group and Lime Creek Formation carbonate platform facies, faunas, and event stratigraphy of northern Iowa, in: Groves, J.R., Walters, J.C. and Day, J., Carbonate Platform Faces and Faunas of the Middle and Upper Devonian Cedar Valley Group and

Lime Creek Formation carbonate platform facies, faunas, and event stratigraphy of northern Iowa. Iowa City, Iowa, Iowa Department of Natural Resources, 15–41.

De Biasi, A.M., Bianchi, C.N. and Morri, C., 2003. Analysis of macrobenthic communities at different taxonomic levels: an example from an estuarine environment in the Ligurian Sea (NE Mediterranean). Estuarine, Coastal and Shelf Science. 58, 99–106.

Fall, L.M. and Olszewski, T.D., 2010. Environmental disruptions influence taxonomic composition of brachiopod paleocommunities in the Middle Permian Bell Canyon Formation (Delaware Basin, West Texas). PALAIOS. 25, 247–259.

Forcino, F.L., Richards, E.J, Leighton, L.R., Chojnacki, N. and Stafford, E.S., 2012. The sensitivity of paleocommunity sampling strategy at different spatiotemporal scales: Palaeogeography, Palaeoclimatology, Palaeoecology. 313, 246–253.

Forcino, F.L., Stafford, E.S. and Leighton, L.R., 2012. Perception of paleocommunities at different taxonomic levels: how low must you go? Palaeogeography, Palaeoclimatology, Palaeoecology. 365, 48-56

Forcino, F.L., Stafford, E.S., Warner, J.J., Webb, A.E., Leighton, L.R., Schneider, C.L., Michlin, T.S., Palazzolo, L.M., Morrow, J.R. and Schellenberg. S.A., 2010. Effects of data categorizations on paleocommunity analysis: a case study from the Pennsylvanian Finis Shale of Texas. PALAIOS. 25, 144–157.

Forcino, F.L. and Leighton, L.R., 2010. Determining the most robust method of measuring disaggregated fossil material: the colonial conundrum. Geological Society of America Abstracts with Programs. 42, 257.

Greffard, M., Saulnier-Talbot, E. and Gregory-Eaves, I., 2011. A comparative analysis of fine versus coarse taxonomic resolution in benthic chironomid community analysis. Ecological Indicators. 11, 1541–1551.

Heim, N.A., 2009. Stability of regional brachiopod diversity structure across the Mississippian/Pennsylvanian boundary. Paleobiology. 35, 393–412.

Holland, S.M. and Patzkowsky, M.E., 2004. Ecosystem structure and stability: Middle Upper Ordovician of Central Kentucky, USA, PALAIOS. 19, 316-331.

Holland, S.M., and Patzkowsky, M.E., 2007, Gradient ecology of a biotic invasion: biofacies of the Type Cincinnatian Series (Upper Ordovician), Cincinnati, Ohio Region, USA. PALAIOS, v, 22, p. 392–407.

Jackson, D.A., 1995. PROTEST: A Procrustean randomization test of community environment concordance. Écoscience. 2, 297–303.

Johnson, J.G., Klapper, G. and Sandberg, C.A., 1985. Devonian eustatic fluctuations in Euramerica. Geological Society of America Bulletin. 96, 567–587.

Kowalewski, M., Gurs, K., Nebelsick, J.H., Oschmann, W., Piller, W.E. and Hoffmeister, A.P., 2002. Multivariate hierarchical analyses of Miocene mollusk assemblages of Europe:

Paleoceanographic, paleoecological, and biostratigraphic implications. GSA Bulletin. 114, 239–256.

Kruskal, J.B., 1964. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychomerika. 29, 1–27.

Lebold, J.G. and Kammer, T.W., 2006. Gradient analysis of faunal distributions associated with rapid transgression and low accommodation space in a Late Pennsylvanian marine embayment: Biofacies of the Ames Member (Glenshaw Formation, Conemaugh Group) in the northern Appalachian Basin. Palaeogeography, Palaeoclimatology, Palaeoecology. 231, 291–314.

Legendre, P. and Legendre, L., 1998. Numerical Ecology. Elsevier, Amsterdam, The Netherlands, 853 p.

Lobza, V., Schieber, J., and Nestell, M., 1994. The Influence of sea level changes and possible pycnocline shifts on benthic communities in the Finis Shale (Virgilian) near Jacksboro, North-Central, Texas. Canadian Society of Petroleum Geologists Memoir. 17, 927–947

Mantel, N., 1967 The detection of disease clustering and a generalized regression approach. Cancer Research. 27, 209–220.

McCune, B. and Grace, J.B., 2002. Analysis of Ecological Communities. MjM Software Design, Gleneden Beach, Oregon, 300 p.

McKinney, F.K. and Hageman, S.J., 2006. Paleozoic to modern ecological shift displayed in the northern Adriatic Sea. Geology. 34, 881-884.

Morris, C., 1986. The community structure of the middle Cambrian Phyllopod Bed (Burgess Shale). Palaeontology. 29, 432–467.

Oksanen, J., Blanchet, F.G., Kindt, R., Legendre, P., O'Hara, R.B., Simpson, G.L., Solymos, P., Stevens, M.H.H. and Wagner, H., 2011. vegan: Community Ecology Package. R package version 1.17-3, http://CRAN.R-project.org/package=vegan.

Olszewski, T.D. and Patzkowski, M.E., 2001. Measuring recurrence of marine biotic gradients: a case study from the Pennsylvanian-Permian Midcontinent. PALAIOS. 16, 444–460.

Olszewski, T.D. and West, R.R., 1997. Influence of transportation and time-averaging in fossil assemblages from the Pennsylvanian of Oklahoma. Lethaia. 30, 315–329.

Patzkowsky, M.E. and Holland, S.E., 1999. Biofacies replacement in a sequence stratigraphic framework: Middle and Upper Ordovician of the Nashville Dome, Tennessee, USA. PALAIOS. 14, 301–323

Peres-Neto, P.R. and Jackson, D.A., 2001. How well do multivariate datasets match? The advantages of a Procrustean superimposition approach over the Mantel Test. Oecologia. 129, 169–178.

Peters, S.E. and Bork, K.B., 1999. Species-abundance models: an ecological approach to inferring paleoenvironment and resolving paleoecological change in the Waldron Shale

(Silurian). PALAIOS. 14, 234–245.

R Development Core Team, 2011. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Redman, C.L., Leighton, L.R., Schellenberg, S.A., Gale, C.N., Neilsen, J.L., Dressler, D.L., and Klinger, M.K., 2007. Influence of spatiotemporal scale on the interpretation of paleocommunity structure: lateral variation in the Imperial Formation of California. PALAIOS. 22, 630–641.

Schneider, C.L., 2003. Community Paleoecology of the Pennsylvanian Winchell Formation, North-Central Texas. Unpublished Ph.D. dissertation, University of Texas, Austin, 306 p.

Schneider, C.L., Forcino, F.L., Barclay, K. and Leighton, L.R., 2012. Significance of paleocommunity analysis for interpreting the ecology and geology of the Devonian Waterways Formation, northeastern Alberta, Canada. Geological Society of America Program with Abstracts. 44.

Staff, G., Powell, E.N., Stanton, R.J., Jr. and Cummins, H., 1985. Biomass: is it a useful tool in paleocommunity reconstruction? Lethaia. 18, 209–232.

Visaggi, C.C. and Ivany, L.C., 2010. The influence of data selection and type of analysis on interpretations of temporal stability in Oligocene fauna of Mississippi. PALAIOS. 25, 769–779.

Warwick, R.M., 1988. Analysis of community attributes of the macrobenthos of Frierfjord/Langesundfjord at taxonomic levels higher than species. Marine Ecology Progress Series. 46, 167–170.

Watkins, R., 1996. Skeletal composition of Silurian benthic marine faunas. PALAIOS. 11, 550–558.

**Table 3-1.** Latitude and longitude of all collections sites along with the number of samples collected at each site.

| Formation | Member | Latitude | Longitude | No. of Samples |
|---|---|---|---|---|
| Cedar Valley | Rapid | N 41.721° | W 91.532° | 15 |
| Waterways | Moberly | N 56.775° | W 111.397° | 2 |
| Waterways | Moberly | N 56.783° | W 111.404° | 2 |
| Waterways | Moberly | N 56.793° | W 111.404° | 3 |
| Waterways | Moberly | N 56.804° | W 111.406° | 1 |
| Waterways | Moberly | N 57.164° | W 111.627° | 1 |
| Waterways | Moberly | N 56.853° | W 111.421° | 1 |
| Waterways | Moberly | N 56.637° | W 111.626° | 1 |
| Waterways | Moberly | N 56.615° | W 111.716° | 1 |
| Waterways | Moberly | N 56.613° | W 111.731° | 2 |
| Waterways | Moberly | N 56.656° | W 111.594° | 1 |
| Waterways | Moberly | N 56.671° | W 111.528° | 1 |
| Waterways | Calumet | N 56.666° | W 110.872° | 1 |
| Waterways | Calumet | N 56.675° | W 110.841° | 5 |

**Table 3-2.** Mantel R-statistics, NMDS axis-one correlation r-values, and PROTEST m2-values along with corresponding p-values for the Rapid Member dataset categorization comparisons. "g-o-f" = goodness-of-fit statistic, "p" = p-value.

| Categorization Comparison | | Mantel R-statistic | Ordination PROTEST $m^2$ | NMDS axis-one r-value |
|---|---|---|---|---|
| Abundance vs. PC (same area) | g-o-f | 0.61 | 0.61 | 0.65 |
| | p | < 0.001 | 0.01 | 0.008 |
| Abundance vs. PC (the entire abundance sample) | g-o-f | 0.71 | 0.78 | 0.78 |
| | p | < 0.001 | < 0.001 | < 0.001 |
| Abundance vs. PC (same area brachiopods only) | g-o-f | 0.78 | 0.84 | 0.63 |
| | p | < 0.001 | < 0.001 | 0.01 |
| Abundance vs. PC (the entire abundance sample brachiopods only) | g-o-f | 0.76 | 0.81 | 0.74 |
| | p | < 0.001 | < 0.001 | 0.002 |
| Abundance vs. PC (standardized) | g-o-f | 0.59 | 0.70 | 0.69 |
| | p | < 0.001 | < 0.001 | 0.005 |
| Abundance All vs. Abundance Brach | g-o-f | 0.83 | 0.82 | 0.59 |
| | p | < 0.001 | < 0.001 | 0.02 |
| Abundance All vs. Abundance Brach (standardized) | g-o-f | 0.85 | 0.88 | 0.76 |
| | p | < 0.001 | < 0.001 | < 0.001 |
| PC All vs. PC Brach | g-o-f | 0.35 | 0.41 | 0.43 |
| | p | 0.09 | 0.18 | 0.11 |
| PC All vs. PC Brach (standardized) | g-o-f | 0.30 | 0.49 | 0.64 |
| | p | 0.02 | 0.04 | 0.01 |

**Table 3-3.** Mantel R-statistics, NMDS axis-one correlation r-values, and PROTEST m2-values along with corresponding p-values for the Waterways Formation dataset categorization comparisons. "g-o-f" = goodness-of-fit statistic, "p" = p-value.

| Categorization Comparison | | Mantel R-statistic | Ordination PROTEST $m^2$ | NMDS axis-one r-value |
|---|---|---|---|---|
| PC vs. Abundance (same area) | g-o-f | 0.73 | 0.73 | 0.90 |
| | p | < 0.001 | < 0.001 | < 0.001 |
| PC vs. Abundance (the entire abundance sample) | g-o-f | 0.74 | 0.73 | 0.90 |
| | p | < 0.001 | < 0.001 | < 0.001 |
| PC vs. Abundance of the entire sample (the entire abundance sample with no crinoids) | g-o-f | 0.85 | 0.92 | 0.93 |
| | p | < 0.001 | < 0.001 | < 0.001 |
| PC vs. Abundance of the entire sample (standardized) | g-o-f | 0.78 | 0.88 | 0.89 |
| | p | < 0.001 | < 0.001 | < 0.001 |

**Table 3-4.** Pearson Product Moment correlation r- and p-values of NMDS axis-one scores for the same comparisons for both units examined in the present study as well as the Finis Shale from Forcino et al. (2010). The r- and p-values for the Finis Shale are slightly different than those in Forcino et al. (2010) because the analyses were conducted using R instead of PCOrd, as originally done in Forcino et al. (2010).

| | Waterways Formation | Rapid Member | Forcino et al. (2010) |
|---|---|---|---|
| All-taxa, abundance vs. biomass proxy | r = 0.90<br>p < 0.001 | r = 0.65<br>p = 0.008 | r = 0.96<br>p < 0.001 |
| All-taxa vs. brachiopods-only, abundance | — | r = 0.59<br>p = 0.02 | r = 0.95<br>p < 0.001 |
| All-taxa vs. brachiopods-only, abundance, biomass proxy | — | r = 0.43<br>p = 0.11 | r = 0.99<br>p < 0.001 |

**Figure 3-1.** General paleogeography of the Middle and Upper Devonian with the sample locations marked (Rp = Rapid Member of Coralville, Iowa and Waterways = Waterways Formations of northern Alberta). The grey areas denote land and the white areas denote water.

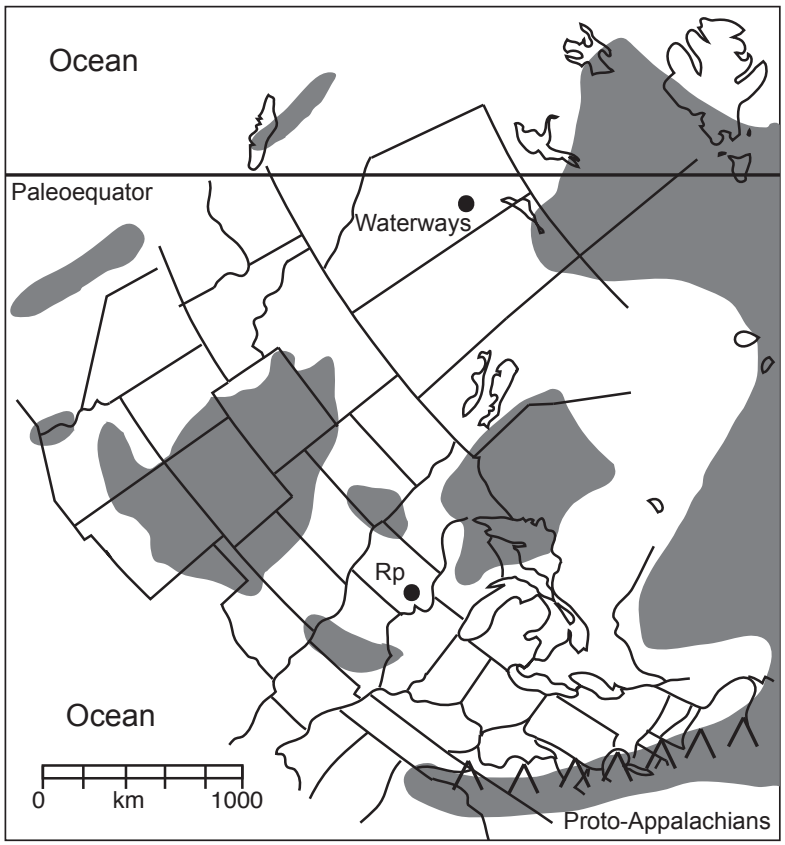


**Figure 3-2.** Middle and Upper Devonian formations of sample locations in Iowa and northern Alberta. The sampled members are shaded in gray.

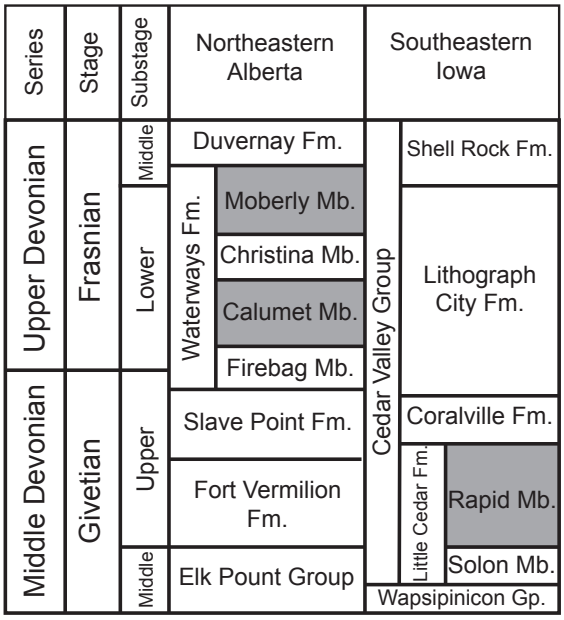


**Figure 3-3.** Non-metric multidimensional scaling ordination plots for the Rapid Member dataset of (A) abundance of all-taxa versus point counts of all-taxa, (B) abundance of all-taxa versus abundance of brachiopod taxa (C) abundance with bryozoans and crinoids culled versus point counts with bryozoans and crinoids culled, and (D) point counts of all-taxa versus point counts of brachiopod taxa. The dashed grey arrows connect points representing the same samples in each categorization. The black points represent abundance-based categorizations, and the gray points represent point count-based categorizations. Circles denote all-taxa, squares denote brachiopods-only, and diamonds denote all-taxa with bryozoans and crinoids culled. All ordinations were run with two dimensions. NMDS stress values are 6.4% for the abundance ordination, 8.7% for the point count ordination, 9.4% for abundance at the MEI level, 8.2% for point counts at the MEI level, 16.3% for abundance without crinoids and bryozoans, 17.2% for point counts without crinoids and bryozoans, 11.7% for abundance of brachiopods, and 15.3% for point counts of brachiopods.

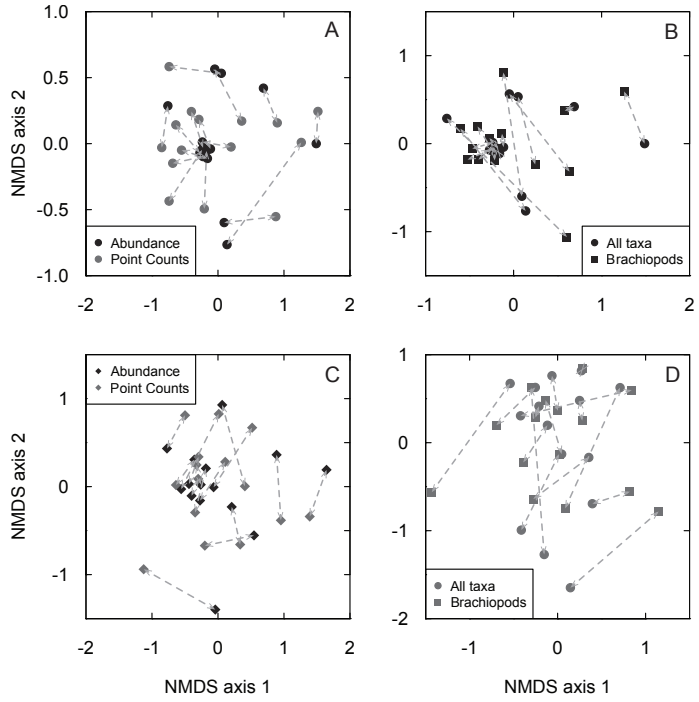


**Figure 3-4.** Non-metric multidimensional scaling ordination plots for the Waterways Formation dataset of (A) abundance of all-taxa versus point counts of all-taxa and (B) abundance with bryozoans and crinoids culled versus point counts with bryozoans and crinoids culled. The dashed grey arrows connect points representing the same samples in each categorization. The black points represent abundance-based categorizations, and the gray points represent point count-based categorizations. Circles denote all-taxa and diamonds denote all-taxa with bryozoans and crinoids culled. All ordinations were run with two dimensions. NMDS stress values are 7.5% for the abundance ordination, 11.3% for the point count ordination, 12.0% for abundance at the MEI level, 10.3% for point counts at the MEI level, 8.4% for abundance without crinoids, and 8.9% for point counts without.

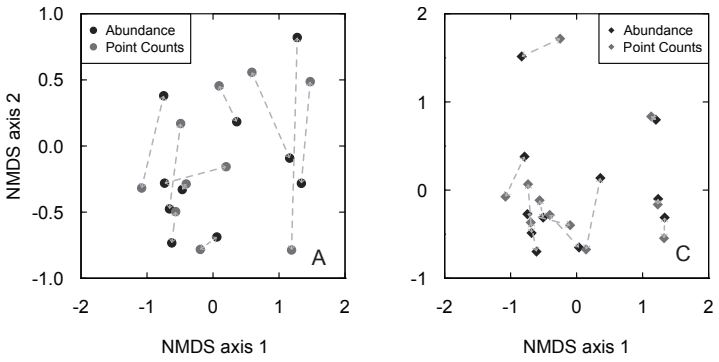


**Figure 3-5.** Three example comparisons of mapped independent variables for categorization comparisons of abundance and points counts of all taxa for the Rapid Member dataset. A, B, and C, are the abundance categorization, and D, E, and F, are the point count categorization. The open gray circles and the closed black circles represent two different groups of independent variables.

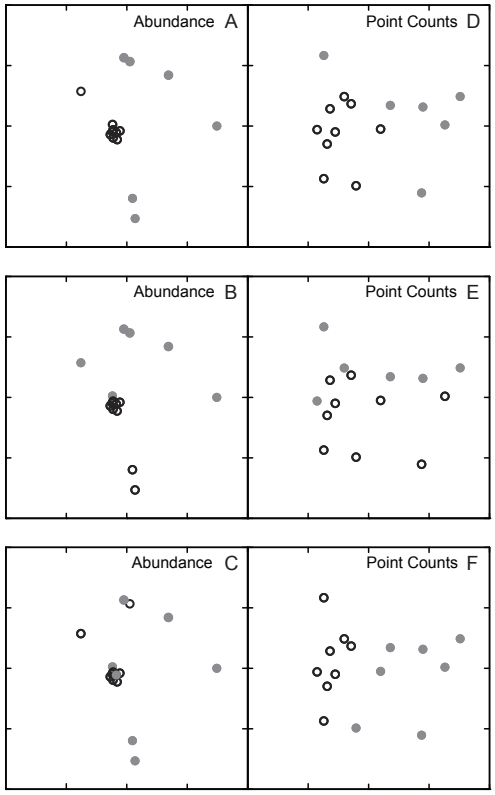


**Figure 3-6.** Three example comparisons of mapped independent variables for categorization comparisons of abundance and points counts of brachiopod-only taxa for the Rapid Member dataset. A, B, and C, are the abundance categorization, and D, E, and F, are the point count categorization. The open gray diamonds and the closed black diamonds represent two different groups of independent variables.

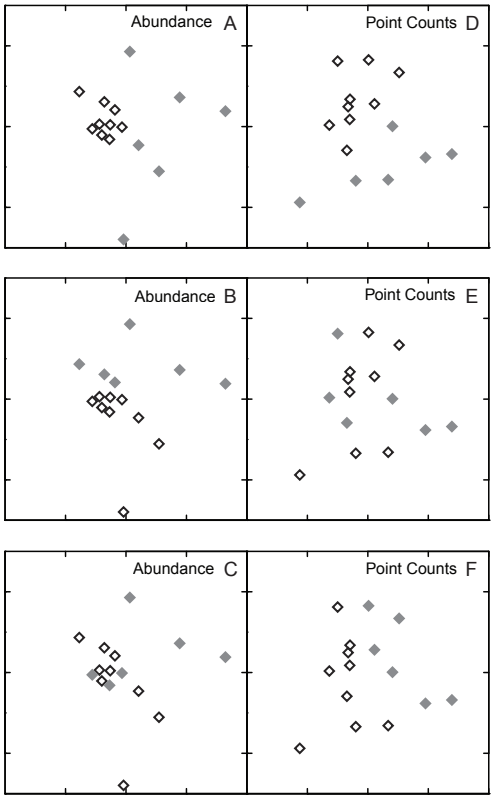


**Figure 3-7.** Three example comparisons of mapped independent variables for categorization comparisons of abundance and points counts of all taxa for the Waterways Formation dataset. A, B, and C, are the abundance categorization, and D, E, and F, are the point count categorization. The open gray circles and the closed black circles represent two different groups of independent variables.

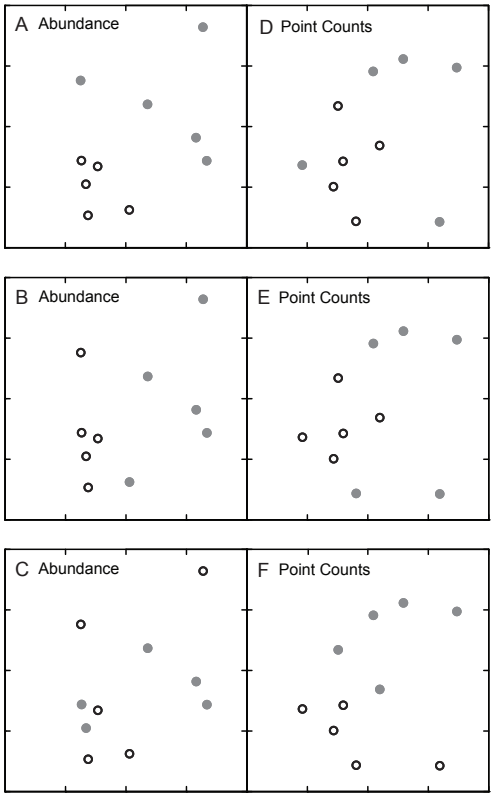


**Figure 3-8.** Three example comparisons of mapped independent variables for categorization comparisons of abundance and points counts of brachiopod-only taxa for the Waterways Formation dataset. A, B, and C, are the abundance categorization, and D, E, and F, are the point count categorization. The open gray diamonds and the closed black diamonds represent two different groups of independent variables.

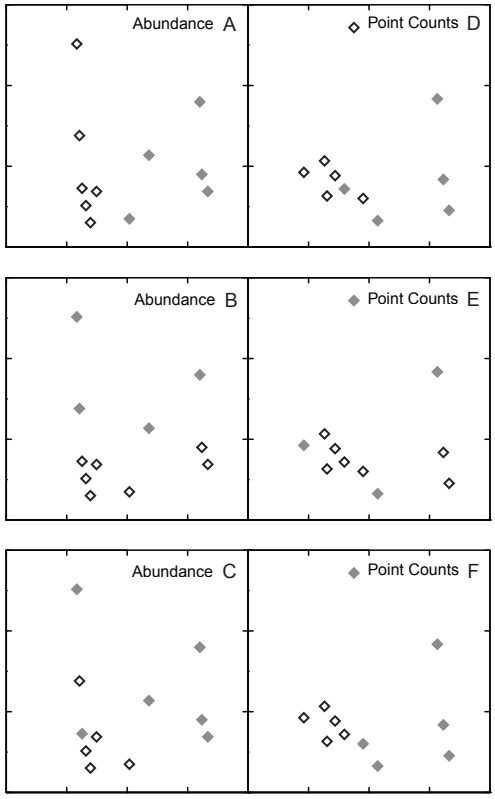


**Figure 3-9.** Three example comparisons of mapped independent variables for categorization comparisons of abundance of all taxa and brachiopod-only taxa for the Rapid Member dataset. A, B, and C, are the abundance categorization, and D, E, and F, are the point count categorization. The open gray points and the closed black points represent two different groups of independent variables.

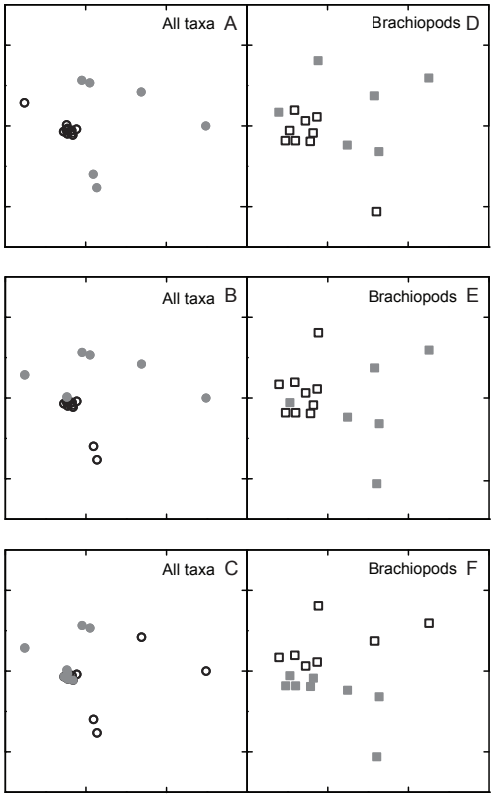


**Figure 3-10.** Three example comparisons of mapped independent variables for categorization comparisons of point counts of all taxa and brachiopod-only taxa for the Rapid Member dataset. A, B, and C, are the abundance categorization, and D, E, and F, are the point count categorization. The open gray points and the closed black points represent two different groups of independent variables.

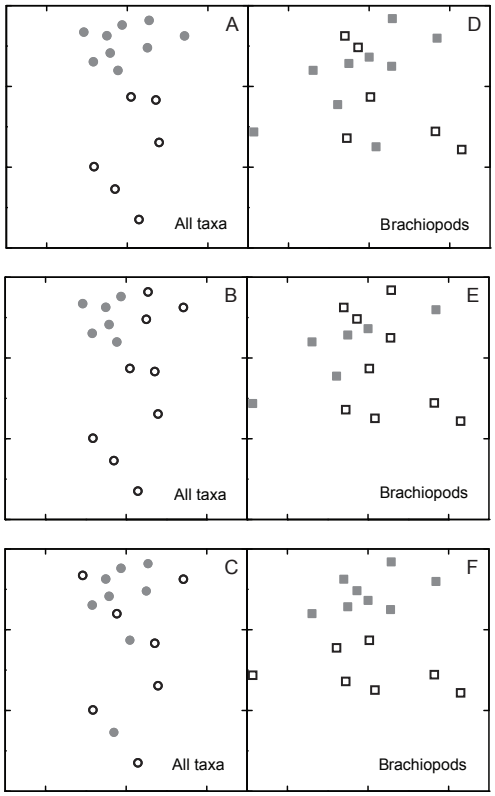
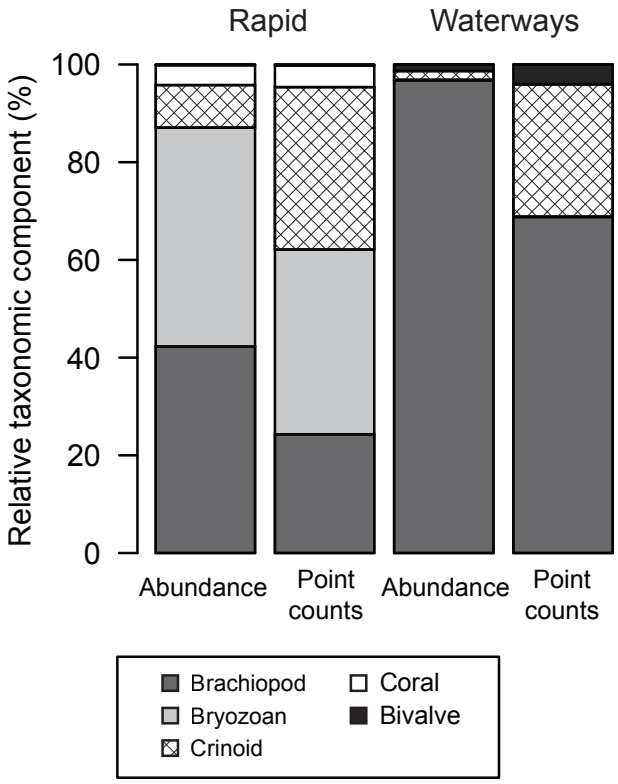


**Figure 3-11.** Relative taxonomic composition of abundance and point count categorizations (both at the generic level) for both the Rapid Member and Waterways Formations datasets.



# Chapter 4. Perception of paleocommunities at different taxonomic levels: how low must you go?[‡]

1. Introduction

Community paleoecology utilizes fossil assemblages to examine the mechanisms of ecological and environmental variation through space and time. Such data provide insight into the processes that structure ecosystems and the causes of ecosystem collapse and extinction (Olszewski and Patzkowsky, 2001; Kowalewski et al., 2002; Bonelli et al., 2006; Redman et al., 2007; Clapham and James, 2008; Leighton and Schneider, 2008; Heim, 2009). In addition, paleocommunity data provide a wealth of information about ecological interactions and environmental tolerances on local scales and help researchers understand how these processes scale up to regional, continental, and global scales (Bambach, 1993; Kowalewski et al., 2002; Bambach et al., 2004; Clapham and Bottjer, 2007). Furthermore, in the face of the rapid tempo of the current biodiversity crisis, conservation science will benefit from increased productivity by paleoecological researchers and quick dissemination of their findings (Bennington and Aronson, 2012). Conducting examinations of how ecosystems were affected by events of rapid warming in the past can lead to predictions regarding how current ecosystems will change (Louys, 2012). This could provide information important to preserving the natural world and counteracting anthropogenic influence.

To obtain meaningful information, researchers must collect samples of fossil assemblages that are sufficiently complete to be accurately representative and large enough to produce statistically robust results (Forcino et al., 2010; Forcino, 2012). Most community paleontological research is resource-intensive, both in terms of travel and in time spent collecting and analyzing data. Therefore, researchers must strike a balance between economy of collection and quantity of data.

A key decision in community paleoecological research is the level of taxonomic identification. Most studies identify specimens to the genus (Olszewski and Patzkowsky, 2001; Forcino et al., 2012) or species (Schneider, 2003). However, some studies use taxonomic families, or non-Linnaean clades or guilds (Kowalewski 2002; Lebold and Kammer, 2006; Forcino et al., 2012). Here, I conducted a meta-analysis of 28 datasets from the Paleobiology Database (PBDB) to determine whether paleocommunity analyses at higher taxonomic levels produce similar results to those analyzed at the genus level. For each dataset, I composed taxon-sample matrices (series of samples containing multiple taxa of varying abundances) at the genus-, family-, order-, and class-levels. Species were not examined, as genus is the lowest taxonomic level included in the PBDB. Genus is also the lowest taxonomic level to which paleocommunity

[‡] *A version of this chapter has been published. Forcino, F.L., Stafford, E.S. and Leighton, L.R. (**2012**) Perception of paleocommunities at different taxonomic levels: how low must you go? Palaeogeography, Palaeoclimatology, Palaeoecology 365:48-56.*

researchers typically identify specimens. I then compared the multivariate paleocommunity results of each of the three higher taxonomic levels (family, order, and class) to the genus-level result. If the higher-level identifications produce the same multivariate results as those at the genus-level, such higher-level classification may be sufficient for community paleoecological research. Simplification of the identification process would conserve resources for researchers. On the other hand, if higher-level identifications do not produce the same multivariate results as the genus-level, then researchers should invest time and effort into lower-level identification in order to produce the most accurate paleocommunity results from the available material.

Identification of fossil specimens to taxonomic levels above genus may provide information as meaningful as information derived from genus- or species-level identification. More closely related organisms tend to be ecologically similar (Darwin, 1859; Cadotte et al., 2008, Cavender-Bares et al., 2009; however, see Cahill et al., 2008 for a counter-argument); two genera of rhynchonellate brachiopods have more in common in terms of natural history, feeding strategy, and environmental preferences than either have with two genera of fish, and vice versa. In addition, identification of specimens at lower taxonomic levels requires much more specialized knowledge than identification to higher levels; a paleocommunity assemblage can contain dozens of genera in several families within multiple phyla, while a typical paleontologist may specialize in only one or two taxonomic groups. When non-specialists attempt to identify specimens to genus or species, identifications may take longer and are more likely to be incorrect. If higher-level classification is sufficient for paleocommunity research, researchers could save significant resources on research projects and reduce inaccurate identification.

Conversely, the taxonomic resolution of paleocommunity data may affect the quality of the paleoecological information derived from the assemblage. While more closely related organisms may tend to be more ecologically similar, even closely related genera can have different ecological functions within a community (Cahill et al., 2008). Lumping these separate genera within a single higher taxon would mask any effect their differing ecologies may have had on the community composition as a whole. The result would be a loss of ability to resolve paleoecological parameters and to detect ecological differences between samples, regions, and formations, degrading the value of the paleocommunity data, and therefore, any conclusions based on those data. Ultimately, the issue is to determine which taxonomic level is more likely to provide the most reliable ecological signal.

Forcino et al. (2010) examined the effect of taxonomic level of identification in paleocommunities from the Finis Shale of Texas, finding a strong correlation between results derived from genus-level identification and results derived from higher-level identifications. Greffard et al. (2011) compared the results of community analyses of chironomid flies (Class Insecta) identified at different taxonomic levels. Although Greffard et al. (2011) concluded that

finer taxonomic resolutions should be employed, the multivariate results of the fine- and coarse-taxonomic resolutions were strongly correlated (PROTEST $m^2 = 0.93$, $p < 0.001$). De Biasi et al. (2003) found similar community results among species-, genus-, and family-level identifications, and a somewhat similar result at the level of taxonomic order. Warwick (1988) found no loss of information (i.e., the same multivariate community results) among species-, genus-, and family-level identifications for five modern marine benthic community datasets.

Alternatively, if a study were to examine a wider range of paleoecological data than those studies described above, it may be discovered that community resolution is lowered when taxa are identified to higher taxonomic levels (e.g., family, order). Combining taxonomic units that may represent different ecological aspects of a community may mask a paleocommunity signal that would be clear when conducting analyses at the genus- or species-level. If the meta-analysis I conduct here finds that higher taxonomic resolution leads to different paleocommunity results, this may reassure researchers that the current paradigm (identifying fossil taxa to genus or species) is necessary to obtain meaningful ecological information from paleocommunities.

2. Methods

Twenty-eight datasets were acquired from the PBDB (Table 4-1). The datasets contained a range of numbers of samples (16 to 162 samples) and generic richness (21 to 244 genera). The datasets represented a wide range of time periods and featured a diversity of taxonomic groups. Each dataset was used to create four taxon-sample matrices—one for each of the taxonomic levels: genus, family, order, and class. The higher-level groupings (family, order, and class) were based on the most up-to-date classifications in the PBDB. Two datasets (Crame, 1981; Budd et al., 1999) contained only one taxonomic order, so no order-level taxon-sample matrices were created from these datasets. Three datasets (Ruzhencev and Bogoslovskaya, 1978; Crame, 1981; Budd et al., 1999) contained only one taxonomic class, so no class-level taxon-sample matrices were created from these datasets.

For each taxonomic-level comparison (family-genus, order-genus, and class-genus) of each dataset, three multivariate statistical comparisons were conducted (Table 4-2):

(1) Using the vegan package in R 2.4 (Oksanen et al., 2011; R Development Core Team, 2011), I performed Mantel Tests of correlation between Bray-Curtis dissimilarity indices (values that quantify the dissimilarity between each object in a taxon-sample matrix). The Mantel Test tests the similarity of two matrices of dissimilarity indices by permuting each of the elements in the dissimilarity matrix 999 times to derive a distribution of correlation values (Mantel, 1967; Fall and Olszewski, 2010). The resulting R-statistic is similar to the Pearson's Product Moment Correlation Coefficient (r); with increasingly similar dissimilarity matrices, the Mantel R-statistic will approach 1. The Mantel Test evaluates the goodness-of-fit of two datasets of non-ordinated data. The remaining two methods evaluate the goodness-of-fit of ordinated data.

(2) I produced non-metric multidimensional scaling (NMDS) ordinations of the samples using the Bray-Curtis dissimilarity index (Clarke and Ainsworth, 1993; Legendre and Legendre, 1998; Bush and Brame, 2010). The NMDS axis-one ordination scores of each taxonomic level were compared to each other pairwise using a Pearson's product-moment correlation.

All NMDS ordinations were run examining the taxonomic distributions among samples. Ordination is an exploratory multivariate visualization tool that allows multidimensional relationships of samples to be examined in fewer dimensions (McCune and Grace, 2002). Because ecological datasets contain samples with taxonomic objects, each with some abundance, ordination is the standard way to visualize the similarities and differences among samples or taxa. Samples that have more similar taxonomic distributions plot closer together in the ordination space.

NMDS ordination iteratively searches for a best-fit solution between the rank dissimilarity indices and the distribution of samples in a low-dimension ordination space. This non-parametric approach is appropriate for community data, which are typically non-normally and non-linearly distributed (Bush and Brame, 2010). The best-fit solution is assessed by the "stress" of the ordination; low stress represents a better NMDS solution (Kruskal, 1964). Stress levels varied from dataset to dataset and from sample size to sample size. NMDS ordinations were run with only two dimensions, which produced stress levels less than 0.20 in all cases.

I used NMDS because it is widely accepted and used among ecologists and paleoecologists, it uses a fixed number of dimensions, and it is non-parametric (Bush and Brame, 2010). I conducted Pearson's Product Moment correlations of NMDS axis-one scores only as means of comparing the ordination results. A more realistic and informative method of comparing NMDS results is the procrustes transformation method (see next paragraph).

(3) Procrustean Randomization Tests (PROTEST) were performed comparing procrustes transformed ordinations of higher- and lower-taxonomic levels (Jackson, 1995; Peres-Neto and Jackson, 2001). The procrustes transformation minimizes the sum-of-squares deviations between the two ordination results through translation, reflection, rotation, and dilation. The PROTEST produced two values: an $m^2$-value, the goodness-of-fit statistic, and a p-value. The residuals between the original solution and the procrustes solution are calculated and produce the $m^2$-value. The $m^2$-value is similar to the r-value resulting from a Pearson's Product Moment Correlation; the closer $m^2$ is to 1, the more similar the two ordinations. The p-value evaluating the significance of the $m^2$-value is calculated by randomly permutating the ordination scores and comparing the randomized values to the original to determine whether they differ significantly from what would be expected by chance.

Goodness-of-fit statistics (i.e., Mantel R-statistics, NMDS axis-one r-values, and PROTEST $m^2$-values) quantify the overall similarity between ordinations, but do not necessarily indicate whether two ordinations would lead a paleoecological researcher to draw the same

conclusions. Many workers interpret ordination results based on details such as identification of clusters of points (samples) in ordination space, often relative to an independent variable. Consistent membership within a cluster, or even identification of a discrete cluster, is not evaluated by any of the multivariate tests described here. The situation is analogous to performing a correlation analysis without subsequently examining the scatterplot. To address this problem, I visually examined each of the results from the procrustes analysis (the translation, reflection, rotation, and dilation portion of the PROTEST) for the 28 genus-family, 26 genus-order, and 25 genus-class comparisons. This not only mimics the process by which many researchers would examine their ordinations for patterns, and so evaluates the importance of details within an ordination plot, but it also provides a means of judging whether the goodness-of-fit statistic values reflect whether the two ordinations would be interpreted in the same way by a paleocommunity researcher. For example, would two ordinations with a PROTEST $m^2$-value of 0.80 be interpreted the same way by paleocommunity researchers? Would PROTEST $m^2$-values of 0.90 or 0.99 indicate sufficient similarity? By examining each ordination comparison individually, I attempted to determine threshold goodness-of-fit statistics for two ordination results to be considered sufficiently similar to be interpreted similarly.

I interpreted two ordination results to be sufficiently similar if fewer than ~10 % of the sample points show a change in position less than 50 % the length of one or both axes. Under these circumstances, I infered that the groupings and patterns of the sample points within the ordinations do not change enough to lead to different paleocommunity interpretations and conclusions (Figures 4-1A and 4-1B). I interpreted two ordinations results as different if greater than 10 % of the sample points show change in position. The sampling units may move considerably along one or both ordination axes. In this case, the changes in the groupings and patterns of the points within the ordinations may lead to different paleocommunity interpretations and conclusions (Figures 4-1C and 4-1D). This assessment was conservative; if I was unsure whether researchers would interpret the ordination results similarly, I labeled the comparisons as different.

In addition to the unmodified analyses described above, I conducted an analysis using the same procedure on manipulated versions of each of the datasets. Within each dataset, all taxa containing fewer than six individuals throughout all samples were culled from the genus-level taxon-sample matrix. Additionally, all samples with fewer than 20 individuals total were culled. The family-, order-, and class-level taxon-sample matrices were then recreated in the same manner as before. Small samples and rare taxa can lead to strong differences in the multivariate analysis of a dataset. By creating culled versions of the datasets, I removed possible outliers from the analyses that may have disproportionately influenced the results. Such culling is a common method in many community paleoecological studies (Redman and Leighton, 2009; Visaggi and Ivany, 2010). Because the results of these analyses were the same as the unmodified and non-

culled dataset analyses, the results and discussion will only address the analyses of the unmodified, non-culled datasets.

To test for autocorrelation (correlation between the different taxonomic levels is due to each higher taxonomic level containing only one representative of the lower taxonomic level), the difference between the number of genera and number of families in each dataset was correlated with three goodness-of-fit statistics (Table 4-3). Twenty-eight values were calculated—the 28 differences between the number of genera and the number of families for each of the 28 datasets. It is possible that the differences in number of taxa between levels is a function of sample sizes (increased size may lead to greater number of genera etc.), but our approach provides a first-order approximation of whether the number of taxa at different levels is driving the goodness-of-fit of the datasets at various taxonomic levels. Pearson's Product Moment correlations were calculated between this set of 28 values and the 28 goodness-of-fit statistics (three times, one for each of the sets of goodness-of-fit statistics). Correlation coefficients closer to one would indicate stronger auto-correlation effects. For completeness, the same correlations were examined using the difference between the numbers of genera and orders as well as between the genera and classes (Table 4-3).

I also conducted Pearson's Product Moment Correlations between the resulting goodness-of-fit statistics and additional dataset variables to determine whether any of these dataset parameters tend to drive similarities between the multivariate results at different taxonomic levels (Table 4-4). The variables I examined were median sample size, median Pielou's J, evenness (Legendre and Legendre, 1998), sample-size-standardized Hurlbert PIE Evenness (Hurlbert, 1971), number of genera, number of families, number of orders, and number of classes.

It is important to note that the basis for taxonomic classification varies among phyla. For example, a family of brachiopods is not automatically equivalent to a family of gastropods in terms of ecological significance, number of genera etc. Some phyla may have different taxonomic levels (e.g. suborders, subfamilies). To determine whether classification system had an effect on how well higher-level results matched with genus-level results, I conducted t-tests on the goodness-of-fit statistics (Table 4-5) for datasets with different dominant phyla: 15 datasets consisting primarily of brachiopods and 10 consisting primarily of mollusks. Three remaining datasets were omitted, two predominantly coral datasets and one predominantly foraminifera dataset. I conducted t-tests between the goodness-of-fit statistics for each brachiopod- and mollusk-dominated dataset for each of the three taxonomic level comparisons (genus-family, genus-class, and genus-order) and for each of the three multivariate comparison methods (Mantel Test, NMDS axis-one correlation, and PROTEST). If the differing classification systems of the two phyla have an effect on the goodness-of-fit statistics, I would expect significant differences between the brachiopod and mollusk groupings.

3. Results

The comparisons of family-genus for the 28 paleocommunity datasets resulted in high goodness-of-fit statistics (Table 4-2; Figure 4-2A). All of the goodness-of-fit statistics for each of the 28 datasets were greater than 0.60 (25 of 28, 89%), with the exception of one Mantel R-statistic and two PROTEST $m^2$-values. Twenty-four Mantel R-statistics, 21 NMDS axis-one r-values, and 16 PROTEST $m^2$-values (61 of 84 goodness-of-fit statistics, 73%) were greater than 0.80. Among the 84 comparisons, 80 (95%) had p-values lower than 0.001, 83 (99%) produced p-values lower than 0.05, and only one (1%) was greater than 0.05. All of the p-values greater than 0.001 resulted from the PROTEST comparisons. Based on the qualitative-visual assessment of the ordination comparisons, 13 genus-family comparisons (13 of 28, 46%) were interpreted as sufficiently similar; the points do not noticeably change position or only a few points change position (Table 4-2). The Mantel R-statistics range from 0.94 to 0.99 for datasets that were interpreted as similar and from 0.62 to 0.94 for those that were interpreted as different. The NMDS r-values range from 0.62 to 0.99 for datasets that were interpreted as similar and from 0.57 to 0.97 for those that were interpreted as different. The PROTEST $m^2$-values range from (0.32) 0.87 to 0.99 for datasets that were interpreted as similar and from 0.35 to 0.92 for those that were interpreted as different. The $m^2$-value of 0.32 is listed as parenthetical because it is a statistical outlier much lower than all of the other values.

The comparisons of order-genus and class-genus for the 28 datasets resulted in moderate to low goodness-of-fit statistics (Table 4-2; Figures 4-2B and 4-2C). Five Mantel R-statistics, eight NMDS axis-one r-values, and five PROTEST $m^2$-values (18 of 78 goodness-of-fit statistics, 23%) were greater than 0.80. Of the 78 comparisons, 66 (85%) produced p-values lower than 0.001. Of the 12 p-values greater than 0.001, one was from the Mantel Tests, five were from the NMDS axis-one correlations, and six were from the PROTEST. Of the 78 comparisons, 75 (96%) produced p-values lower than 0.05; one NMDS axis-one correlation and two PROTEST comparisons were greater than 0.05. Based on the qualitative-visual assessment of the ordination comparisons, 3 genus-family comparisons were interpreted as the same (Table 4-2).

Among class-genus comparisons, two Mantel R-statistics, four NMDS axis-one r-values, and two PROTEST $m^2$-values (8 of the 75 goodness-of-fit statistics, 11%) were greater than 0.80. Of the 75 comparisons, 57 (73%) produced p-values lower than 0.001. Of the 18 p-values greater than 0.001, one was from the Mantel Tests, nine were from the NMDS axis-one correlations, and eight were from the PROTEST. Of the 75 comparisons, 72 (96%) produced p-values lower than 0.05; one NMDS axis-one correlation and two PROTEST comparisons were greater than 0.05. Based on the qualitative-visual assessment of the ordination comparisons, all class-genus comparisons were interpreted as different (Table 4-2).

Using Pearson's Product Moment Correlations of the three goodness-of-fit statistics

versus the difference between the numbers of genera and families (as well as the differences between the numbers of genera and orders and between the numbers of genera and classes), I found no significant correlation for any of the three goodness-of-fit statistics for any of the three different taxonomic level comparisons (Table 4-3), indicating no autocorrelation between number of genera and of families. I also ran correlations between the goodness-of-fit statistics and several other dataset variables, including the number of samples, two measures of taxonomic evenness (one sample size dependent, one sample size independent), and the absolute numbers of genera, of families, of orders, and of classes (Table 4-4). Out of 54 (3 tests× 3 taxonomic comparisons× 6 variables) total comparisons, only four correlations were significant at a $p < 0.05$. The r-statistics associated with these four significant correlations were weak, ranging from 0.4-0.5. None of the correlations were significant at $p < 0.001$.

I conducted t-tests between the goodness-of-fit statistics for each brachiopod- and mollusk-dominated dataset for each of the three taxonomic level comparisons (genus-family, genus-class, and genus-order) and for each of the three multivariate comparison methods (Mantel Test, NMDS axis-one correlation, and PROTEST). All of the nine t-tests (one for each combination of taxonomic level comparison and multivariate comparison method) produced non-significant differences between the brachiopod and mollusk groupings (lowest $p = 0.33$; Table 4-5).

4. Discussion

Because the genus- and family-level identifications led to the same paleocommunity results approximately half of the time (46%), it is possible that paleoecological researchers may acquire accurate and meaningful results using family-level identifications when the alternative choice is genus-level identification (Table 4-2; Figure 4-2). However, many differences arose between genus- and family-level ordinations within many of the datasets; based on our visual inspection, I considered 15 of the 28 genus-family comparisons to be different (Table 4-2; Figures 4-1 and 4-2). In multiple cases, samples moved dramatically in position from one ordination to the next, enough to change membership within clusters of samples, even when the two associated data matrices were evaluated as significantly similar by the multivariate tests. Thus, I recommend using caution when using taxonomic levels higher than genus. To ensure consistent and accurate paleocommunity interpretations and conclusions, genus-level identification of specimens is probably the best choice.

I was conservative in judging whether two ordinations were the same or different; other researchers may consider some of the 15 different comparisons to be similar, as seven of these 15 are borderline-similar. However, these data cast doubt on the similarity between genus- and family-level identifications and suggest that family is not a sufficient substitute for genus identifications in paleocommunity research.

Even if the family-level is sufficient for most paleocommunity research, it is likely that identification to the genus, or even species, level would result in additional ecological insight. Particularly, if the primary goal of a study is to determine the maximum diversity of an ecosystem, or to examine communities where closely-related genera or species are known to differ ecologically, identification of specimens to genus or species would almost certainly yield better information. Therefore, researchers should consider the goals of individual projects to determine whether the time and resources spent identifying specimens to low taxonomic levels may be better spent collecting more samples or embarking on additional projects. Furthermore, some paleontologists argue that conducting analyses using any taxonomic level higher than species leads to uninformative results. In cases where multiple species within a genus are present in a community, these species may differ enough ecologically to consider the community distinct from one containing only one species of the genus. Combining the two species into the single genus may reduce the accuracy of the paleocommunity signal. Regardless, most paleocommunity research is conducted at the genus level, and it is not uncommon for most genera to include only a single species within a given paleocommunity.

The results derived from the order- and class-level identifications clearly and consistently did not produce the same results as genus-level identifications (Table 4-2; Figures 4-1 and 4-2). This suggests that higher-level taxonomic identifications (i.e., order and class) do not reliably reproduce the multivariate community results found with genus-level identification.

Auto-correlation is one possible explanation for the strong similarities between the genus- and family-level results of half of the datasets: if each family in an assemblage is represented by only a single genus (i.e., each genus has its own separate family), the data matrices at each level are essentially identical, despite the different names applied to the specimens within the matrix. If the family-genus comparisons in our study are auto-correlated, our conclusion (that family-level identification sometimes yields results as accurate as genus-level identification) may not extend to datasets with multiple genera within single families. To test for autocorrelation, the difference between the number of genera and number of families in each dataset was correlated with three goodness-of-fit statistics. Pearson's Product Moment Correlations were calculated between this set of 28 values and the 28 goodness-of-fit statistics (three times, one for each of the sets of goodness-of-fit statistics). I found no significant correlation for any of the three goodness-of-fit statistics for any of the three different taxonomic level comparisons (Table 4-3). Based on theses analyses, auto-correlation does not appear to be driving our results.

It is also possible that other community variables may make certain datasets more prone to high or low similarity between genus-level and higher-level results. For example, datasets consisting of samples with highly even taxonomic distributions may behave differently than datasets consisting of samples with low taxonomic evenness. I ran correlations between the

goodness-of-fit statistics and various dataset variables (number of samples, two measures of taxonomic evenness, number of genera, number of families, number of orders, and number of classes). Out of 63 total correlations, only four were significant at $p < 0.05$ and zero were significant at $p < 0.001$ (Table 4-4). No correlations within a given dataset variable were significant for all three goodness-of-fit tests, and significant correlations did not consistently appear for any taxonomic-level comparison. Furthermore, the r-statistics associated with the four significant correlations were weak, ranging from 0.4 to 0.5. Thus, the data do not support any influence from sample size, evenness, number of genera, number of families, number of orders, or number of classes on the amount of similarity between genus-level and higher-level multivariate results.

Another possible reason that genus- and family-level identifications produced similar results for some cases (especially compared to the order- and class-level identifications) is that the genera within each family led ecologically similar lifestyles (Darwin, 1859; Cadotte et al., 2008, Cavender-Bares et al., 2009; however, see Cahill et al., 2008 for a counter-argument). In other words, most genera within a family are functionally similar. While this is not the case for all genera, it may be a reasonable generalization for many paleocommunities. Taxa that perform similar ecological functions may not need to be individually identified for multivariate paleocommunity analyses. The differences between samples may be strongly influenced by the primary ecological functions of the taxa within the communities. Therefore, a similar multivariate result may be produced even when these genera are grouped together as families or higher clades. This might explain the similar paleocommunity results between genus- and higher -level identifications in previous work (Lebold and Kammer, 2002; Forcino et al., 2010). Order- and class-level results may differ from genus-level results because of fundamentally different ecological processes operating at higher taxonomic levels. It may be that paleocommunity signals produced using higher-level identifications obscure the signal of the processes that control the generic (or family) composition of communities.

If I could determine the circumstances under which families act as adequate proxies for genera, this knowledge would be extremely useful for future paleocommunity research. However, given the present results, it would seem that if there is a single common cause, it is not any of the variables (e.g., evenness, number of samples) explored herein.

Although genus-level identification may be required within the range of parameters represented by the 28 datasets I obtained and analyzed from the PBDB, there may be datasets for which family- or higher-level taxonomic identifications are more appropriate. In other words, the required range of paleocommunity variables and the subsequent required taxonomic level might depend on the particular research questions being explored. The use of finer taxonomic resolution (genus or species) may add noise to an otherwise precise result. For example, if examining continental-scale processes using paleocommunities, identifying taxa to the generic

level may only add variation that represents local scale processes. In fact, it may not be possible to evaluate broad-scale processes using the genus-level because individual genera may not have a sufficient environmental range to compare across such large geographic scales, while families or orders would be more likely to be geographically widespread.

It is important to note that the basis for taxonomic classification varies among phyla; an order within one phylum is not necessarily equivalent (in terms of number of genera, morphological and genetic disparity, evolutionary history, etc.) to an order within another phylum. To determine whether differing classification systems had an effect on how well higher-level results matched with genus-level results, I conducted t-tests between the goodness-of-fit statistics for each brachiopod- and mollusk-dominated dataset for each of the three taxonomic level comparisons (genus-family, genus-class, and genus-order) and for each of the three multivariate comparison methods (Mantel Test, NMDS axis-one correlation, and PROTEST). If classification system has an effect on the goodness-of-fit statistics, I would expect significant differences between the brachiopod and mollusk groupings. All of the nine t-tests (one for each combination of taxonomic level comparison and multivariate comparison method) produced non-significant differences between the brachiopod and mollusk groupings (lowest $p = 0.33$; Table 4-5). Thus, differences between the classification systems did not affect the comparisons of paleocommunity results. Note, however, that this comparison is only between brachiopod and mollusk classification; both phyla are similar in that they are solitary, benthic, macroinvertebrates with external shells. Phyla that tend to be colonial (e.g., corals and bryozoans), microscopic or very small (e.g., foraminiferans), or composed of numerous skeletal pieces (e.g., crinoids, vertebrates) may have significantly different classification systems. Many of the datasets used few datasets dominated by organisms other than brachiopods or mollusks, I was not able to examine any effects the classification systems of other phyla may have on comparisons among taxonomic levels. However, the datasets used here were not chosen at random; many of the datasets used by paleocommunity researchers primarily consist of brachiopods and mollusks.

Further, there may be differences between higher and lower taxonomic groupings of colonial organisms, such as corals or bryozoans compared to higher and lower groupings of solitary brachiopods and mollusks. There are more ecological and morphological differences between corals and brachiopods or mollusks than between brachiopods and mollusks. So, colonial animals may be a case where higher and lower taxonomic levels may produce more similar results. For example, colonial morphology is often very plastic, and it is often difficult to identify colonial animals to species or genus. So, if colonial animals did produce the same multivariate results between higher and lower taxonomic levels, this information could be even more useful than when examining solitary taxa. However, it is also possible that higher taxa better capture the colonial morphology, which presumably is an important ecological

indicator. In either case, there is not enough information in the present study to provide any robust insight into this issue. I only had two datasets dominated by corals; the PBDB does not contain many community datasets of corals along with abundance information. Future research on the subject of the appropriate taxonomic level for identifying colonial organisms could be extremely useful for paleocommunity studies.

## 5. Conclusion

High goodness-of-fit statistics resulted between comparisons (using three different statistical comparison methods) of genus- and family-level taxonomic identifications for 28 paleocommunity datasets. However, 15 of the 28 genus-family comparisons were determined to produce different paleocommunity results based on qualitative-visual comparisons. Thus, family-level identification of specimens can lead to the same paleocommunity conclusions as genus-level identification, but inconsistencies generate enough uncertainty that paleocommunity research would benefit from genus-level identification of specimens. Due to the moderate-to-low goodness-of-fit statistics between genus-order and genus-class comparisons of paleocommunities, as well as the clear differences found in the qualitative-visual comparisons, order and class did not reliably reproduce genus-level results. Thus, while family-level identifications may be sufficient at times for studies employing multivariate statistical methods to compare paleocommunities that would otherwise use the genus level, order- and class-level identifications were never sufficient within the meta-analysis. If the similarity that was found between some of the genus- and family-level comparisons is due to the ecological similarity among the genera within a family, higher clade-level identification may be sufficient if justified by ecological similarity among clade members. The idea of ecological similarity is supported by additional studies that have found strong correlations between the paleocommunity results of genus- and higher clade-level identification of specimens (Lebold and Kammer, 2002; Forcino et al., 2010). Autocorrelation, evenness, and number of samples, genera, families, orders, or classes did not affect the level of similarity between any of the comparisons of any of the three taxonomic levels. Furthermore, possible differences in classification system between brachiopods and mollusks did not produce differences in the level of similarity between any of the comparisons of any of the three taxonomic levels.

## 6. Acknowledgements

This research was supported by NSF-EAR0746072 grant to Leighton, Schellenberg, Morrow, and Schneider. Thank you to Chris Schneider, Amelinda Webb, and Darrin Molinaro for input on results and critiques of an early draft of the manuscript. I would like to especially thank Tom Olszewski and one anonymous reviewer for their comments and suggestions that significantly improved this manuscript.

## 7. Bibliography

Ahmad, F., 1998. Taxonomy and palaeoecology of the benthic macroinvertebrate fauna from the

Middle Jurassic of northwestern Jordan. PhD Thesis, University of Wurzburg.

Bambach, R.K, Knoll, A.H. and Well, J.B., 2004. Origination, extinction, and mass depletions of marine diversity. Paleobiology. 30, 522-542.

Bennington, J.B. and Aronson, M.F.J., 2012. Reconciling scale in paleontological and neontological data: dimensions of time, space, and taxonomy, in: Louys, J. (Ed.), Paleontology in Ecology and Conservation. Springer-Verlag, Berlin, pp. 39-67.

Bonelli Jr., J.R., Brett, C.E., Miller, A.I. and Bennington, J.B., 2006. Testing for faunal stability across a regional biotic transition: quantifying stasis and variation among recurring coral-rich biofacies in the Middle Devonian Appalachian Basin. Paleobiology. 32, 20–37.

Budd, A.F., Johnson, K.G., Stemann, T.A., Tompkins, B.H., 1999. Pliocene to Pleistocene reef coral assemblages in the Limon Group of Costa Rica. Bulletins of American Paleontology 357, 119–158.

Bulinski, K.V., 2007. Analysis of sample-level properties along a paleoenvironmental gradient: the behavior of evenness as a function of sample size. Palaeogeography, Palaeoclimatology, Palaeoecology 253, 490–508.

Bush, A.M. and Brame, R.I., 2010. Multiple paleoecological controls on the composition of marine fossil assemblages from the Frasnian (Late Devonian) of Virginia, with a comparison of ordination methods. Paleobiology. 36, 573–591.

Cadotte, M.W., Cardinale, B.J. and Oaklet, T.H., 2008. Evolutionary history and the effect of biodiversity on plant productivity. Proceedings of the National Academy of Sciences. 105, 17012-17017.

Cahill, J.F., Kembel, S.W., Lamb, E.G. and Keddy, P.A., 2008. Does phylogenetic relatedness influence the strength of competition among vascular plants? Perspectives in Plant Ecology, Evolution and Systematics. 10, 41-50.

Cavender-Bares, J., Kozak, K.H., Fine, P.V.A. and Kembel, S.W., 2009. The merging of community ecology and phylogenetic biology. Ecology Letters. 12, 693-715.

Clapham, M.E., Bottjer, D.J., 2007. Permian marine paleoecology and its implications for large-scale decoupling of brachiopod and bivalve abundance and diversity during the Lopingian (Late Permian). Palaeogeography, Palaeoclimatology, Palaeoecology 249, 283–301.

Clapham, M.E. and James, N.P., 2008. Paleoecology of Early-Middle Permian marine communities in eastern Australia: response to global climate change in the aftermath of the late Paleozoic ice age. PALAIOS. 23, 738-750.

Clarke, K.R. and Ainsworth, M., 1993. A method and linking multivariate community structure to environmental variables. Marine Ecology. 92, 205–219.

Crame, J.A., 1981. Ecological stratification in the Pleistocene coral reefs of the Kenya coast.

Palaeontology 24, 609–646.

Darwin, C. 1859. The Origin of Species. John Murray, London.

De Biasi, A.M., Bianchi, C.N. and Morri, C., 2003. Analysis of macrobenthic communities at different taxonomic levels: an example from an estuarine environment in the Ligurian Sea (NE Mediterranean). Estuarine, Coastal and Shelf Science. 58, 99-106.

Dominici, S., Kowalke, T., 2007. Depositional dynamics and the record of ecosystem stability: early Eocene faunal gradients in the Pyrenean foreland, Spain. Palaios 22, 268–284.

Fall, L.M., Olszewski, T.D., 2010. Environmental disruptions influence taxonomic composition of brachiopod paleocommunities in the Middle Permian Bell Canyon Formation (Delaware Basin, West Texas). Palaios 25, 247–259

Forcino, F.L., 2012. Multivariate assessment of the required sample size for paleocommunity research. Palaeogeography, Palaeoclimatology, Palaeoecology. 315, 134-141.

Forcino, F.L., Richards, E.J, Leighton, L.R., Chojnacki, N. and Stafford, E.S., 2012. The sensitivity of paleocommunity sampling strategy at different spatiotemporal scales. Palaeogeography, Palaeoclimatology, Palaeoecology. 313, 246-253.

Forcino, F.L., Stafford, E.S., Warner, J.J., Webb, A.E., Leighton, L.R., Schneider, C.L., Michlin, T.S., Palazzolo, L.M., Morrow, J.R. and Schellenberg. S.A., 2010. Effects of data categorizations on paleocommunity analysis: a case study from the Pennsylvanian Finis Shale of Texas. PALAIOS. 25, 144–157.

Fürsich, F.T., Wendt, J., 1997. Biostratinomy and palaeoecology of the Cassian Formation (Triassic) of the Southern Alps. Palaeogeography, Palaeoclimatology, Palaeoecology 22, 257–323.

Gahr, M.E., 2002. Palökologie benthischer Faunen aus dem Unter-Toarc SW-Europas. Ph.D. Thesis, Universität Würzburg.

Greffard, M., Saulnier-Talbot, E. and Gregory-Eaves, I., 2011. A comparative analysis of fine versus coarse taxonomic resolution in benthic chironomid community analysis. Ecological Indicators. 11, 1541-1551.

Heim, N.A., 2009. Stability of regional brachiopod diversity structure across the Mississippian/ Pennsylvanian boundary. Paleobiology 35, 393–412.

Hogler, J.A., 1992. Community structure, paleoecology, and depositional environments within the Upper Triassic Luning Formation of Nevada. PhD Thesis, University of California, Berkeley.

Holland, S.M., Patzkowsky, M.E., 2004. Ecosystem structure and stability: middle Upper Ordovician of central Kentucky, USA. Palaios 19, 316–331.

Holzapfel, S., 1998. Palökologie benthischer Faunengemeinschaften und Taxonomie der Bivalven im Jura von Südtunesien. Beringeria — Würzburger geowissenschaftliche Mitteilungen.

Hurlbert, S.H., 1971. The nonconcept of species diversity: a critique and alternative parameters. Ecology 52, 577-586.

Jackson, D.A., 1995. PROTEST: A Procrustean randomization test of community environment concordance. Écoscience 2, 297–303.

Kowalewski, M., Gurs, K., Nebelsick, J.H., Oschmann, W., Piller, W.E., Hoffmeister, A.P., 2002. Multivariate hierarchical analyses of Miocene mollusk assemblages of Europe: paleoceanographic, paleoecological, and biostratigraphic implications. GSA Bulletin 114, 239–256.

Kruskal, J.B., 1964. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychomerika 29, 1–27.

Lebold, J.G. and Kammer, T.W., 2006. Gradient analysis of faunal distributions associated with rapid transgression and low accommodation space in a Late Pennsylvanian marine embayment: Biofacies of the Ames Member (Glenshaw Formation, Conemaugh Group) in the northern Appalachian Basin. Palaeogeography, Palaeoclimatology, Palaeoecology. 231, 291-314.

Legendre, P. and Legendre, L., 1998. Numerical Ecology. Elsevier, Amsterdam, The Netherlands, 853 p.

Leighton, L.R, and Schneider, C.L., 2008.Taxon characteristics that promote survivorship through the Permian-Triassic interval: transition from Paleozoic to the Mesozoic brachiopod fauna. Paleobiology. 34, 65-79

Louys, J., 2012. Paleontology in ecology and conservation: an introduction, in: Louys, J. (Ed.), Paleontology in Ecology and Conservation. Springer-Verlag, Berlin, pp. 1-7.

Mantel, N., 1967. The detection of disease clustering and a generalized regression approach. Cancer Research 27, 209–220.

McCune, B., Grace, J.B., 2002. Analysis of Ecological Communities. MjM Software Design, Gleneden Beach, Oregon.

Oksanen, J., Blanchet, F.G., Kindt, R., Legendre, P., O'Hara, R.B., Simpson, G.L., Solymos, P., Stevens, M.H.H. and Wagner, H., 2011. vegan: Community Ecology Package. R package version 1.17-3, http://CRAN.R-project.org/package=vegan.

Olszewski, T.D., Patzkowsky, M.E., 2001. Measuring recurrence of marine biotic gradients: a case study from the Pennsylvanian–Permian Midcontinent. Palaios 16, 444–460.

Patzkowsky, M.E., 1995. Gradient analysis of Middle Ordovician brachiopod biofacies: Biostratigraphic, biogeographic, and Macroevolutionary Implications. Palaios 10, 154–179.

Peres-Neto, P.R. and Jackson, D.A., 2001. How well do multivariate datasets match? The advantages of a Procrustean superimposition approach over the Mantel Test. Oecologia. 129, 169–178.

Peters, S.E., 2003. Evenness, Richness and the Cambrian-Paleozoic Faunal Transition in North America: An Assemblage-Level Perspective. Ph.D. Thesis, The University of Chicago.

Popov, L.E., Cocks, L.R.M., Nikitin, I.F., 2002. Upper Ordovician brachiopods from the Anderken Formation, Kazakhstan: their ecology and systematics. Bulletin of the Natural History Museum, Geology Series 58, 13–79.

Redman, C.L., Leighton, L.R., Schellenberg, S.A., Gale, C.N., Neilsen, J.L., Dressler, D.L., and Klinger, M.K., 2007. Influence of spatiotemporal scale on the interpretation of paleocommunity structure: lateral variation in the Imperial Formation of California. PALAIOS. 22, 630-641.

Redman, C.L. and Leighton, L.R., 2009. Multivariate faunal analyses of the Turonian Bissekty Formation: variation in the degree of marine influence in temporally and spatially averaged fossil assemblages. PALAIOS, 24, 18-26.

Ruzhencev, V.E. and Bogoslovskaya, M.F., 1978. Namyurskii etap v evolyutsii ammonoidyey - Namurian stage, evolution of ammonoids. *Truydi paleontologicheskovo instituta,* 338 pp.

Schneider, C.L., 2003. Community paleoecology of the Pennsylvanian Winchell Formation, north-central Texas. Ph.D. Thesis, The University of Texas.

Schubert, J.K., Bottjer, D.J., 1995. Aftermath of the Permian-Triassic mass extinction event: Paleoecology of Lower Triassic carbonates in the western USA. Palaeogeography, Palaeoclimatology, Palaeoecology 116, 1–39.

Scott, R.W., 1970. Paleoecology and paleontology of the Lower Cretaceous Kiowa Formation, Kansas. The University of Kansas Paleontological Contributions 52, 1–94.

Sohl, N.F. and Koch, C.F., 1984. Upper Cretaceous (Maastrichtian) larger invertebrate Fossils from the Haustator bilira Assemblage Zone in the west gulf Coastal Plain. USGS Open File Report. 687, 1-282.

Stilwell, J.D., Zinsmeister, W.J., Oleinik, A.E., 2004. Early Paleocene mollusks of Antarctica: Systematics, paleoecology and paleobiogeographic significance. Bulletins of American Paleontology 367, 1–89.

Taylor, P.D., Smith, A.B. and McGowan, A.J., 2006. Unpublished field notes and sections.

Tomasovych, A., 2006. Brachiopod and bivalve ecology in the Late Triassic (Alps, Austria): onshore–offshore replacements caused by variations in sediment and nutrient supply. Palaios 21, 344–368.

Visaggi, C.C. and Ivany, L.C., 2010. The influence of data selection and type of analysis on interpretations of temporal stability in Oligocene fauna of Mississippi. PALAIOS 25, 769-779.

Warwick, R.M., 1988. Analysis of community attributes of the macrobenthos of Frierfjord/Langesundfjord at taxonomic levels higher than species. Marine Ecology Progress Series. 46, 167-170.

Waterhouse, J.B., 1987. Late Palaeozoic mollusca and correlations from the south-east Bowen Basin, east Australia. Palaeontographica Abteilung A 198, 129–333.

Wienrich, G., 1997. Die Fauna des marinen Miozäns von Kevelaer (Niederrhein). Backhuys Publishers, Leiden, 1-187

Williams, A., 1974. Ordovician Brachiopoda from the Shelve District, Shropshire. Bulletin of the British Museum of Natural History (Geology). 11, 1-158

Williams, A., Lockley, M.G., Hurst, J.M., 1981. Benthic palaeocommunities represented in the Flairfach Group and coeval Ordovician successions of Wales. Palaeontology 24, 661–694.

**Table 4-1.** Listing of all 28 datasets used for the taxonomic-level comparisons, with attributes. Brach = brachiopod, Cam = Cambrian, Ord = Ordovician, Miss = Mississippian, Penn = Pennsylvanian.

| Reference | PBDB ref # | Time Periods | Primary Lithology | Primary Group | # of Genera | # of Families | # of Orders | # of Classes | # of Samples | Median Sample Size |
|---|---|---|---|---|---|---|---|---|---|---|
| Scott, 1970 | 52 | Cretaceous | Sandstone | Mollusk | 53 | 41 | 23 | 10 | 74 | 27 |
| Williams, 1974 | 11552 | Ord | Shale | Brach | 62 | 37 | 11 | 4 | 25 | 6 |
| Fürsich and Wendt, 1977 | 23366 | Triassic | Limestone | Mollusk | 161 | 101 | 37 | 10 | 51 | 61 |
| Ruzhencev and Bogoslovskaya, 1978 | 19354 | Penn | Limestone | Mollusk | 42 | 12 | 3 | NA | 66 | 95 |
| Crame. 1981 | 18189 | Neogene | Limestone | Coral | 28 | 11 | NA | NA | 38 | 9 |
| Williams, et al., 1981 | 162 | Ord | Sandstone | Brach | 29 | 26 | 13 | 7 | 75 | 78 |
| Sohl and Koch, 1984 | 282 | Cretaceous | Sandstone | Mollusk | 216 | 112 | 35 | 9 | 104 | 53 |
| Waterhouse, 1987 | 11597 | Permian | Sandstone | Brach | 130 | 61 | 31 | 11 | 162 | 47 |
| Hogler, 1992 | 8769 | Triassic | Limestone | Mollusk | 31 | 25 | 14 | 7 | 85 | 37 |
| Watkins, 1994 | 4161 | Silurian | Limestone | Brach | 30 | 24 | 15 | 9 | 19 | 99 |
| Patzkowsky, 1995 | 401 | Ord | Limestone | Brach | 25 | 18 | 7 | 3 | 43 | 47 |
| Schubert and Bottjer, 1995 | 8833 | Triassic | Limestone | Mollusk | 27 | 21 | 14 | 6 | 34 | 69 |
| Wienrich, 1997 | 18176 | Paleogene | Sandstone | Foram | 255 | 146 | 28 | 9 | 62 | 219 |
| Ahmad, 1998 | 9529 | Jurassic | Limestone | Brach | 50 | 30 | 15 | 5 | 52 | 575 |
| Holzapfel, 1998 | 6119 | Jurassic | Limestone | Mollusk | 107 | 61 | 27 | 8 | 96 | 66 |
| Budd et al., 1999 | 24129 | Neogene | Limestone | Coral | 31 | 12 | NA | NA | 112 | 15 |
| Gahr, 2002 | 7068 | Jurassic | Limestone | Brach | 103 | 62 | 25 | 7 | 144 | 67 |
| Popov et al., 2002 | 26964 | Ord | Sandstone | Brach | 87 | 50 | 22 | 12 | 41 | 21 |
| Schneider, 2003 | 11551 | Penn | Shale | Brach | 34 | 26 | 15 | 6 | 72 | 40 |
| Peters, 2003 | 13544 | Cam-Ord | Limestone | Brach | 122 | 72 | 28 | 14 | 30 | 166 |
| Holland and Patzkowsky, 2004 | 9838 | Ord | Limestone | Brach | 46 | 31 | 23 | 13 | 94 | 45 |
| Stilwell et al., 2004 | 19080 | Paleogene | Sandstone | Mollusk | 70 | 40 | 16 | 2 | 38 | 10 |
| Clapham, 2006 | 13431 | Permian | Limestone | Brach | 148 | 89 | 31 | 13 | 16 | 43 |
| Taylor et al., 2006 | 18152 | Triassic | Limestone | Brach | 42 | 26 | 19 | 9 | 58 | 18 |
| Tomasovych, 2006 | 17562 | Triassic | Limestone | Mollusk | 21 | 18 | 10 | 2 | 20 | 47 |
| Bulinski, 2007 | 24916 | Ord | Limestone | Brach | 24 | 19 | 12 | 8 | 139 | 27 |
| Dominici and Kowalke, 2007 | 25442 | Paleogene | Shale | Mollusk | 68 | 43 | 17 | 5 | 98 | 103 |
| Heim, 2009 | 26838 | Miss-Penn | Limestone | Brach | 45 | 31 | 10 | 3 | 61 | 28 |

**Table 4-2.** Listing of all 28 datasets used for the taxonomic-level comparisons, with goodness-of-fit statistics and p-values. All p-values listed as 0.001 are < 0.001. The portions highlighted in gray are the comparisons I interpreted as the same based on visual comparison of the procrustes transformed ordination results.

| Reference | Genus-Family | | | | | | Genus-Order | | | | | | Genus-Class | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | R | p | r | p | $m^2$ | p | R | p | r | p | $m^2$ | p | R | p | r | p | $m^2$ |
| Scott, 1970 | 0.93 | 0.001 | 0.80 | 0.001 | 0.67 | 0.001 | 0.79 | 0.001 | 0.49 | 0.001 | 0.46 | 0.001 | 0.55 | 0.001 | 0.47 | 0.001 | 0.30 |
| Williams, 1974 | 0.85 | 0.001 | 0.85 | 0.001 | 0.78 | 0.001 | 0.54 | 0.001 | 0.70 | 0.001 | 0.59 | 0.001 | 0.52 | 0.001 | 0.68 | 0.001 | 0.59 |
| Fürsich and Wendt, 1977 | 0.99 | 0.001 | 0.97 | 0.001 | 0.95 | 0.001 | 0.77 | 0.001 | 0.51 | 0.001 | 0.42 | 0.001 | 0.52 | 0.001 | 0.38 | 0.005 | 0.45 |
| Ruzhencev and Bogoslovskaya, 1978 | 0.70 | 0.001 | 0.57 | 0.001 | 0.35 | 0.003 | 0.34 | 0.001 | 0.15 | 0.24 | 0.12 | 0.43 | NA | NA | NA | NA | NA |
| Crame. 1981 | 0.81 | 0.001 | 0.73 | 0.001 | 0.62 | 0.001 | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA |
| Williams, et al., 1981 | 0.99 | 0.001 | 0.99 | 0.001 | 0.99 | 0.001 | 0.78 | 0.001 | 0.79 | 0.001 | 0.78 | 0.001 | 0.77 | 0.001 | 0.79 | 0.001 | 0.77 |
| Sohl and Koch, 1984 | 0.92 | 0.001 | 0.85 | 0.001 | 0.72 | 0.001 | 0.69 | 0.001 | 0.67 | 0.001 | 0.61 | 0.001 | 0.39 | 0.001 | 0.48 | 0.001 | 0.38 |
| Waterhouse, 1987 | 0.73 | 0.001 | 0.80 | 0.001 | 0.70 | 0.01 | 0.49 | 0.001 | 0.71 | 0.001 | 0.64 | 0.002 | 0.45 | 0.001 | 0.40 | 0.001 | 0.34 |
| Hogler, 1992 | 0.99 | 0.001 | 0.97 | 0.001 | 0.92 | 0.001 | 0.77 | 0.001 | 0.69 | 0.001 | 0.63 | 0.001 | 0.60 | 0.001 | 0.71 | 0.001 | 0.62 |
| Watkins, 1994 | 0.62 | 0.001 | 0.79 | 0.001 | 0.71 | 0.001 | 0.33 | 0.002 | 0.65 | 0.002 | 0.52 | 0.005 | 0.30 | 0.002 | 0.67 | 0.002 | 0.60 |
| Patzkowsky, 1995 | 0.86 | 0.001 | 0.93 | 0.001 | 0.85 | 0.001 | 0.51 | 0.001 | 0.59 | 0.001 | 0.49 | 0.001 | 0.44 | 0.001 | 0.62 | 0.001 | 0.48 |
| Schubert and Bottjer, 1995 | 0.98 | 0.001 | 0.97 | 0.001 | 0.98 | 0.001 | 0.89 | 0.001 | 0.86 | 0.001 | 0.87 | 0.001 | 0.62 | 0.001 | 0.52 | 0.002 | 0.75 |
| Wienrich, 1997 | 0.97 | 0.001 | 0.99 | 0.001 | 0.98 | 0.001 | 0.90 | 0.001 | 0.99 | 0.001 | 0.92 | 0.001 | 0.90 | 0.001 | 0.99 | 0.001 | 0.92 |
| Ahmad, 1998 | 0.99 | 0.001 | 0.94 | 0.001 | 0.92 | 0.001 | 0.73 | 0.001 | 0.83 | 0.001 | 0.82 | 0.001 | 0.65 | 0.001 | 0.78 | 0.001 | 0.78 |
| Holzapfel, 1998 | 0.94 | 0.001 | 0.79 | 0.001 | 0.87 | 0.001 | 0.68 | 0.001 | 0.74 | 0.001 | 0.63 | 0.001 | 0.37 | 0.001 | 0.24 | 0.02 | 0.21 |
| Budd et al., 1999 | 0.95 | 0.001 | 0.98 | 0.001 | 0.98 | 0.001 | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA |
| Gahr. 2002 | 0.92 | 0.001 | 0.97 | 0.001 | 0.92 | 0.001 | 0.75 | 0.001 | 0.33 | 0.001 | 0.39 | 0.001 | 0.43 | 0.001 | 0.30 | 0.001 | 0.31 |
| Popov et al., 2002 | 0.65 | 0.001 | 0.88 | 0.001 | 0.83 | 0.001 | 0.43 | 0.001 | 0.74 | 0.001 | 0.68 | 0.001 | 0.39 | 0.001 | 0.59 | 0.001 | 0.62 |
| Schneider, 2003 | 0.98 | 0.001 | 0.97 | 0.001 | 0.99 | 0.001 | 0.88 | 0.001 | 0.93 | 0.001 | 0.92 | 0.001 | 0.74 | 0.001 | 0.89 | 0.001 | 0.90 |
| Peters, 2003 | 0.95 | 0.001 | 0.62 | 0.001 | 0.32 | 0.115 | 0.68 | 0.001 | 0.45 | 0.012 | 0.44 | 0.006 | 0.56 | 0.001 | 0.45 | 0.01 | 0.40 |
| Holland and Patzkowsky, 2004 | 0.94 | 0.001 | 0.92 | 0.001 | 0.74 | 0.001 | 0.72 | 0.001 | 0.33 | 0.002 | 0.31 | 0.001 | 0.45 | 0.001 | 0.10 | 0.37 | 0.21 |
| Stilwell et al., 2004 | 0.99 | 0.001 | 0.90 | 0.001 | 0.92 | 0.001 | 0.79 | 0.001 | 0.86 | 0.001 | 0.80 | 0.001 | 0.64 | 0.001 | 0.41 | 0.01 | 0.50 |
| Clapham, 2006 | 0.89 | 0.001 | 0.79 | 0.001 | 0.60 | 0.012 | 0.74 | 0.001 | 0.78 | 0.001 | 0.49 | 0.06 | 0.64 | 0.001 | 0.67 | 0.004 | 0.45 |
| Taylor et al., 2006 | 0.96 | 0.001 | 0.94 | 0.001 | 0.89 | 0.001 | 0.93 | 0.001 | 0.93 | 0.001 | 0.79 | 0.001 | 0.71 | 0.001 | 0.74 | 0.001 | 0.75 |
| Tomasovych, 2006 | 0.80 | 0.001 | 0.95 | 0.001 | 0.90 | 0.001 | 0.57 | 0.001 | 0.73 | 0.001 | 0.63 | 0.001 | 0.44 | 0.001 | 0.67 | 0.001 | 0.59 |
| Bulinski, 2007 | 0.95 | 0.001 | 0.97 | 0.001 | 0.96 | 0.001 | 0.85 | 0.001 | 0.89 | 0.001 | 0.59 | 0.001 | 0.81 | 0.001 | 0.95 | 0.001 | 0.73 |
| Dominici and Kowalke, 2007 | 0.84 | 0.001 | 0.95 | 0.001 | 0.61 | 0.001 | 0.60 | 0.001 | 0.87 | 0.001 | 0.66 | 0.001 | 0.57 | 0.001 | 0.85 | 0.001 | 0.68 |
| Heim, 2009 | 0.80 | 0.001 | 0.60 | 0.001 | 0.60 | 0.001 | 0.49 | 0.001 | 0.64 | 0.15 | 0.33 | 0.002 | 0.45 | 0.001 | 0.22 | 0.09 | 0.31 |

**Table 4-3.** Correlation coefficients and p-values for Pearson's Product Moment Correlations for the auto-correlation tests. The difference between the number of taxonomic groups was calculated, then correlated with the corresponding goodness-of-fit statistics.

| | Mantel Test | | NMDS axis-one | | PROTEST | |
|---|---|---|---|---|---|---|
| Genus to Family | r = 0.04 | p = 0.83 | r = 0.11 | p = 0.56 | r = 0.14 | p = 0.47 |
| Genus to Order | r = 0.06 | p = 0.78 | r = 0.10 | p = 0.63 | r = 0.14 | p = 0.51 |
| Genus to Class | r = 0.13 | p = 0.52 | r = 0.06 | p = 0.78 | r = 0.07 | p = 0.72 |

**Table 4-4.** Correlation coefficients and p-values for Pearson's Product Moment Correlations for the correlations between the goodness-of-fit statistics and six additional variables. Fam = comparisons of genus to family, Order = comparisons of genus to order, and Class = comparisons of genus to class.

| | | Mantel Test | | NMDS axis-one | | PROTEST | |
|---|---|---|---|---|---|---|---|
| | | r-statistic | p-value | r-statistic | p-value | r-statistic | p-value |
| Fam | Number of Samples | 0.11 | 0.58 | 0.21 | 0.28 | 0.08 | 0.69 |
| Order | | 0.05 | 0.80 | 0.10 | 0.63 | 0.05 | 0.79 |
| Class | | 0.13 | 0.55 | 0.50 | 0.01 | 0.25 | 0.22 |
| Fam | Pielou's J Evenness | 0.13 | 0.49 | 0.02 | 0.91 | 0.08 | 0.68 |
| Order | | 0.21 | 0.31 | 0.07 | 0.75 | 0.10 | 0.64 |
| Class | | 0.21 | 0.32 | 0.18 | 0.38 | 0.10 | 0.63 |
| Fam | Hurlbert PIE Evenness | 0.09 | 0.63 | 0.06 | 0.76 | 0.06 | 0.78 |
| Order | | 0.32 | 0.12 | 0.01 | 0.97 | 0.06 | 0.78 |
| Class | | 0.21 | 0.30 | 0.18 | 0.37 | 0.17 | 0.41 |
| Fam | Number of Genera | 0.13 | 0.52 | 0.27 | 0.17 | 0.08 | 0.69 |
| Order | | 0.16 | 0.45 | 0.03 | 0.90 | 0.07 | 0.74 |
| Class | | 0.06 | 0.79 | 0.31 | 0.13 | 0.15 | 0.48 |
| Fam | Number of Families | 0.19 | 0.32 | 0.23 | 0.25 | 0.02 | 0.92 |
| Order | | 0.23 | 0.25 | 0.03 | 0.90 | 0.09 | 0.65 |
| Class | | 0.10 | 0.64 | 0.24 | 0.25 | 0.13 | 0.55 |
| Fam | Number of Orders | 0.21 | 0.29 | 0.32 | 0.11 | 0.03 | 0.89 |
| Order | | 0.28 | 0.17 | 0.19 | 0.37 | 0.02 | 0.91 |
| Class | | 0.12 | 0.56 | 0.46 | 0.02 | 0.40 | 0.05 |
| Fam | Number of Classes | 0.07 | 0.75 | 0.21 | 0.31 | 0.40 | 0.05 |
| Order | | 0.07 | 0.75 | 0.29 | 0.15 | 0.28 | 0.17 |
| Class | | 0.04 | 0.85 | 0.16 | 0.44 | 0.26 | 0.21 |

**Table 4-5.** T-test results for comparisons of goodness-of-fit statistics between predominantly mollusk datasets and predominantly brachiopod datasets.

| | Mantel Test | | NMDS axis-one | | PROTEST | |
|---|---|---|---|---|---|---|
| Genus to Family | t = 0.767 | p = 0.45 | t = 0.071 | p = 0.95 | t = 0.005 | p = 0.99 |
| Genus to Class | t = 0.455 | p = 0.65 | t = 0.370 | p = 0.72 | t = 0.030 | p = 0.98 |
| Genus to Order | t = 0.602 | p = 0.55 | t = 0.999 | p = 0.33 | t = 0.659 | p = 0.52 |

**Figure 4-1.** Examples of ordination comparisons that were interpreted to be the same (A and B), and different (C and D). The black points represent the ordination result based on genus-level identifications, and the gray points represent the ordination result based on higher-level identifications. The dashed lines connect corresponding sampling units.

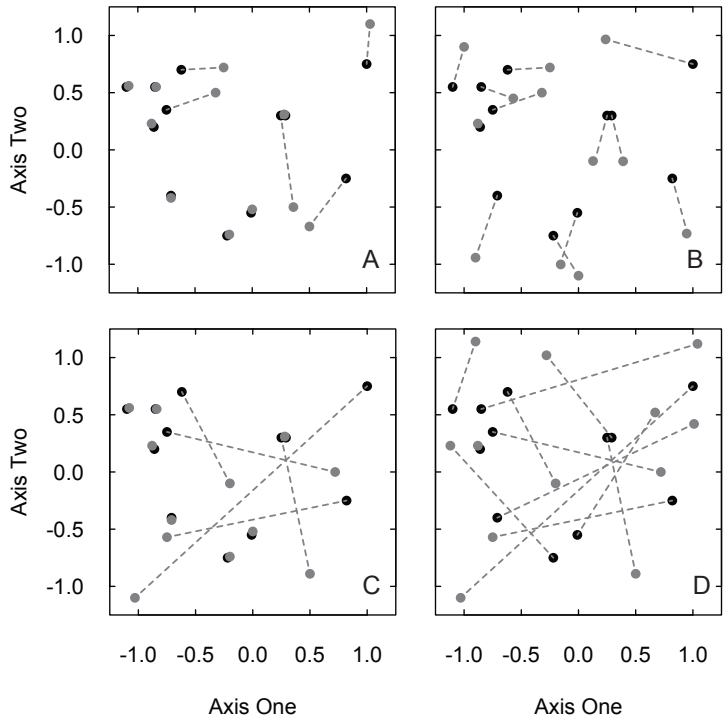
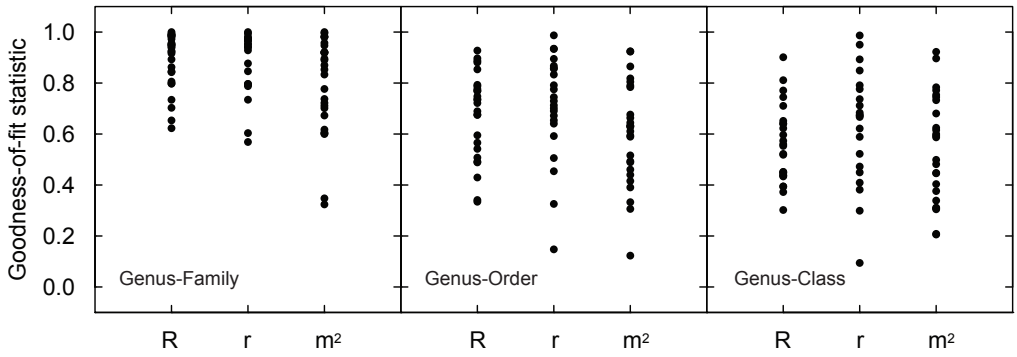


**Figure 4-2.** All goodness-of-fit statistics for the (A) family-genus, (B) order-genus, and (C) class-genus comparisons of the 28 datasets for (from left to right within each plot) Mantel Test R-statistics, NMDS axis-one correlation coefficients, and PROTEST $m^2$-values.



# Chapter 5. Multivariate assessment of the required sample size for community paleoecological research[§]

1. Introduction

Community paleoecology utilizes complete fossil assemblages to study spatiotemporal ecological and environmental variation, leading to insights into the processes that structure ecosystems (Olszewski and Patzkowsky, 2001; Kowalewski et al., 2002; Bonelli et al., 2006; Clapham et al., 2006; Clapham and James, 2008; Heim, 2009). To be able to determine the most accurate information about paleocommunities, paleoecologists must collect statistically representative samples (Hayek and Buzas, 1997; Bennington and Rutherford, 1999; Visaggi and Ivany, 2010). An important component of the statistical viability of a paleocommunity sample is the number of individuals per sample (i.e., sample size). Without a sample size large enough to be representative of the complete fossil assemblage, erroneous ecological conclusions can be reached.

Paleocommunity research is labor- and time-intensive, as it requires expertise in multiple taxonomic groups and sub-disciplines of geology to develop a stratigraphic and biological framework (Kowalewski et al., 2002; Forcino et al., 2010a). Therefore, it is important to not over-sample, because an increasingly larger sample size will eventually result in diminishing returns in terms of improving any pattern revealed by the fossil data. A paleoecologist's time may be better-spent acquiring additional samples to supplement other types of statistical power (Bennington, 2003; Zambito et al., 2008).

Previous studies estimated sample size requirements to be 300 or more individuals per sample based on the probability of acquiring species that makeup 1% of a sample with 95% confidence (Phleger, 1960; Chang, 1967; Patterson and Fishbein, 1989; Fatela and Taborda, 2002). However, examining probability does not take into account relationships between sampling units or taxa (e.g., the interaction of taxa within a community or evolutionary changes of taxa through time). Studying a range of datasets that have been used to study paleocommunities is one manner to account for these relationships.

Rarefaction has been used to determine whether a sample is sufficiently large to represent taxonomic richness (Heck et al., 1975; Caron and Jackson, 2008). If a rarefaction curve plateaus above a certain sample size, it is determined that the sample size is sufficient. Rarefaction is also used to standardize for inconsistent sample sizes (Raup, 1975; Miller and Foote, 1996). In these cases, a smaller sample size is simulated by drawing random samples from the original population.

---

[§] *A version of this chapter has been published. Forcino, F.L. (**2012**) Multivariate assessment of the required sample size for paleocommunity research. Palaeogeography, Palaeoclimatology, Palaeoecology 315:134-141.*

Although some researchers have used the above studies or subsampling procedures to establish a sample size for a particular research project, currently there is no standard protocol for determining the required sample size for paleocommunity studies. For example, individual samples in three typical community studies ranged from 7 to 213 with a median of 28 individuals (Heim, 2009), 113 to 1059 with a median of 43 individuals (Clapham, 2006), and 110 to 883 with a median of 323 individuals (Forcino et al., 2010a). One of the primary goals herein is to develop a recommended sampling protocol for paleocommunity research. If an evidence-based recommended protocol is established, the data from various studies will be more compatible. This will aid meta-studies (Wagner, 2006; Alroy et al., 2008), which are becoming more common due to the use of the Paleobiology Database (PBDB) and other databases (e.g., Integrated Ocean Drilling Program).

The majority of paleocommunity research examines patterns of communities using multivariate statistical techniques such as cluster analyses and ordination (e.g., Olszewski and Patzkowsky, 2001; Holland, 2005; Tomasovych, 2006; Redman et al., 2007; Bonelli and Patzkowsky, 2008; Clapham and James, 2008; Forcino et al., 2010a). Although using theoretical considerations (e.g., probability, rarefaction) to estimate sufficient sample size is a valid method, comparing community datasets at different sample sizes using multivariate techniques is a practical means of ascertaining the effect of sample size on paleocommunity research. Here, I subsampled 30 datasets from the Paleobiology Database and the literature from a wide range of time periods, taxonomic groups, scales, lithologies, numbers of samples (13–124), and numbers of taxa (21–167), and initial median sample sizes (9–2441). I compared each subsample to its corresponding complete dataset in order to determine the smallest sample sizes that produced the same overall pattern of relationships between component samples.

## 2. Methods

### 2.1 Dataset Acquisition

A typical dataset used for multivariate analysis of paleocommunities consists of a number of taxa contained within a series of samples, distributed either stratigraphically or spatially. A dataset used for multivariate analysis is organized as a taxon-sample matrix. Taxon-sample matrices were obtained from the PBDB, from the author's collections, and from the literature (Table 5-1). These matrices all contain marine invertebrate assemblages and represent a broad range of time periods, taxonomic groups, geographic scales, and lithologies and contain various numbers of samples and taxa.

### 2.2 Subsampling Protocol

Using R 2.11.1 (R Development Core Team, 2010), each sample within each taxon-sample matrix was randomly subsampled without replacement to five proportional sizes: 50%, 25%, 10%, 5%, and 2.5% of the total number of individuals. For example, if sample A contained 24 individuals and sample B contained 14 individuals, the 50% subsampled taxon-sample matrix

would consist of 12 individuals from sample A and 7 individuals from sample B. Because the total number of individuals from each sample is randomly selected from a pool of all individuals from the full taxonomic distribution of each sample, the abundance of each taxon is not necessarily 50% of its original value; rather, its abundance has a hypergeometric error distribution around the expected value.

2.2.1 Multivariate Analysis

For each of the subsampled proportions (e.g. 50%, 25%, etc.) of each taxon-sample matrix (dataset), 1000 subsampled matrices were constructed for a total of 5000 subsampled matrices for each dataset. Each of the 5000 subsampled matrices were statistically compared to the original 100% taxon-sample matrix using three multivariate methods:

(1) Using the vegan package in R (Oksanen et al., 2010), Mantel Tests of correlation were performed between the Bray-Curtis dissimilarity indices (values that quantify the dissimilarity between each object in a taxon-sample matrix; McCune and Grace, 2002) of each of the 5000 taxon-sample matrices and the corresponding Bray-Curtis dissimilarity indices of the 100% matrix. No transformations or standardizations were performed on any of the datasets (in order to be consistent throughout the subsampling process). The Mantel Test tests the similarity of two matrices of dissimilarity indices by permuting each of the elements in the dissimilarity matrix 999 times to derive a distribution of correlation values (Mantel, 1967; Fall and Olszewski, 2010). The resulting R-statistic is similar to the Pearson's Product Moment Correlation Coefficient (r); with increasingly similar dissimilarity matrices, the Mantel R-statistic will approach 1.

(2) For each of the 5000 dissimilarity matrices, non-metric multidimensional scaling (NMDS) ordinations of the samples were performed using the Bray-Curtis dissimilarity index (Clarke and Ainsworth, 1993; Bush and Brame, 2010). For each of the 5000 NMDS ordinations performed for each dataset, the axis-one score was correlated with the axis-one score of the NMDS ordination of the 100% dataset.

Here, all ordinations were run examining the taxonomic distributions among samples. Ordination is an exploratory multivariate visualization tool that allows multidimensional relationships of samples to be examined in a low number of dimensions (McCune and Grace, 2002). Because ecological datasets contain samples with taxonomic objects, each with some abundance, ordination is the standard way to visualize the similarities and differences among samples or taxa. Samples that have more similar taxonomic distributions plot closer together in the ordination space.

NMDS ordination iteratively searches for a best-fit solution is reached between the rank dissimilarity indices and the distribution of samples in a low dimension ordination space. This non-parametric approach is appropriate for community data, which are typically non-normally and non-linearly distributed (Bush and Brame, 2010). The best-fit solution is assessed by the

stress of the ordination; low stress represents a better NMDS solution (Kruskal, 1964). Stress levels varied from dataset to dataset and from sample size to sample size. Most stress levels were less than 0.20, but were sometimes higher because all NMDS ordinations were run with only two dimensions in order to be consistent and conservative.

A false "best" placement of the samples in ordination space can occur that is different than the actual best result if a local minimum of stress is found. The NMDS algorithm may mistakenly interpret a relative low stress as the absolutely lowest stress. These types of variation in NMDS results, that might not be an issue when using a different ordination method or a greater number of dimensions, can cause type II error: improper rejection of the hypothesis that there is no significant difference between the subsampled axis-one score and the 100% dataset axis-one score. Thus, for the purposes of the present study, using NMDS axis-one scores, rather than other ordination method axis-one scores, is a conservative approach, in that NMDS may be more likely to fail to identify a proportional subset as similar to its 100% dataset. Other ordination methods that are not random may have more of a chance for repeating the same result based on the user-chosen parameters. However, all ordination methods have advantages and disadvantages. NMDS was used here because it is a widely accepted and used ordination method among ecologists and paleoecologists, it uses a fixed number of dimensions, and it is non-parametric (Bush and Brame, 2010).

To assess the effect a dissimilarity measure of 1 would have on the results of NMDS comparisons, five datasets were subsampled using the flexible shortest path adjustment (FSPA) of ecological distances used by Bush and Brame (2010). These results were the same as those not using the FSPA method. Thus, there was no detrimental effect to the subsampling protocol when there is a dissimilarity measure of 1 in a dataset.

(3) A procrustes transformation was performed between each of the NMDS comparisons (i.e., complete dataset versus each of the subsampled dataset NMDS) and the similarity was assessed using Procrustean Randomization Test (PROTEST; Jackson, 1995, Peres-Neto and Jackson, 2001). The procrustes analysis minimizes the sum-of-square deviations between the two ordination results through translation, reflection rotation, and dilation. The PROTEST compares randomized values to the procrustes result and determines if it differs significantly from what would be expected by chance. The resulting $m^2$-value is similar to the r-value resulting from the Pearson's Product Moment Correlation; the closer $m^2$ is to 1, the more similar the two ordinations.

2.2.2 Graphical Representation

For each of the five subsample proportions for each dataset, the mean Mantel test R-statistic, the mean correlation coefficient, and the mean $m^2$-value of the 1000 dissimilarity matrices, NMDS axis-one scores, and NMDS ordinations, respectively, were plotted versus the corresponding median sample size (Figure 5-1). Sample sizes that produce the same

paleocommunity results will all have the same general R-statistics, r-values, and $m^2$-values. Lower R-statistics, r-values, and $m^2$-values represent a sample size that is too small to produce the same paleocommunity result as that of larger sizes. This area on the plot will display lower correlation statistics. If there is a breaking point on the curve where there is a clear difference between the higher and lower correlation statistics, this may be interpreted as a sample size threshold at which subsample results are representative of the 100% sample. The mean correlation statistic of each set of subsamples is used to produce the figures because these results were normally distributed, while median sample sizes are used because sample size distributions for all of the complete taxon-sample matrices were right-skewed.

Any independent variables (e.g., lithology or time period) that produce consistently different correlation statistics represent important information for determining the appropriate samples size when conducting studies where these variables are known. To determine any potential lithologic, temporal, richness, or initial median sample size trends in R-statistics, r-values, and $m^2$-values, the corresponding time period and lithology of each dataset were mapped onto the plots of all 30 complete datasets comparing median sample sizes versus mean R-statistics, r-values, and $m^2$-values (Table 5-2; Figure 5-2). If there were a consistent separation between correlation statistics based on lithology, time periods, richness bins (i.e., high and low), or complete dataset median sample size bins (i.e., large and small) this would be evidence that lithology or time period affects the sample size required for multivariate paleocommunity research. To test for any differences between the bins of each of the four variables, t-tests were performed between the correlation statistics for all median sample sizes greater than 50 (Table 5-3).

2.3 Diversity Metrics

In addition to multivariate methods of examining how community samples vary through time or space, diversity metrics are often used as an additional means of quantifying differences between samples and communities (Bennington, 2003; Heim, 2009). Richness (S), Pielou's Evenness (J), and Shannon's Enthalpy (H) were calculated for each sample within each of the 1000 subsample matrices for all five subsample sizes for each of the 30 datasets (Magurran, 2004). Richness is the number of taxa in a sample, evenness is a measure of how evenly distributed the taxa are within each sample, and Shannon's H is a measure of diversity that takes into account both richness and evenness. For each of the 30 datasets, the mean S, J, and H were calculated for each sample within all 1000 subsampled datasets for all five proportions. These mean values were then correlated with the corresponding S, J, and H of each of the corresponding samples in the 100% matrix. The mean correlation coefficients of S, J, and H for the five proportions of each dataset were plotted against the corresponding median sample sizes to determine the effect of sample size on S, J, and H (Figure 5-3).

2.4 Devonian Rapid Member of Iowa

To assess the degree to which gregariousness (i.e., congeneric clustering in a sample) may influence the results of the subsampling protocol used herein, a paleocommunity dataset collected for this study was included in the analysis for two purposes: (1) it was used as one of the 30 datasets included in the subsampling program, and (2) it was used as a case study examining how samples collected at various sample sizes in the field compared to the computer generated simulated subsampling results. These samples from the Givetian Rapid Member of the Little Cedar Formation at the Devonian Fossil Gorge in Coralville, Iowa (N 41° 43.3', W 91° 31.9') demonstrated strong gregariousness among some fossil taxa within Rapid Member paleocommunities (Leighton and Schneider, 2004). The Devonian Fossil Gorge consists of 300 horizontal meters of multiple terraced limestone platforms alternating between argillaceous micrite to biosparite and massive floatstone (Bunker and Witzke, 1992). Samples were collected from the 15 least weathered platform surfaces; the total sampled section from the base of the Rapid Member was 8.7 m.

At each surface, a sampling location was selected randomly and six samples were collected at six different sample sizes (5, 10, 20, 50, 100, and 200 individuals). Subsampling was nested; individuals contained within each smaller sample size were included as a portion of the next larger sample size. Only one of each sample size was collected. For example, only one sample size containing 10 individuals was collected as opposed to 20 making up the complete 200-individual sample size. Each of the six datasets produced by each of the different sample sizes were run through the same multivariate analyses as described above for use in the subsampling protocol. Multivariate results of the smaller sample sizes were correlated with the multivariate results of the dataset containing 200 individuals per sample.

This produced two sets of Mantel R-statistics, two sets of NMDS axis-one r-values, and two sets of PROTEST $m^2$-values. One set resulted from the mean R-statistics, r-values, and $m^2$-values produced by the simulated subsampling of the complete 200-specimen dataset. The other set resulted from a statistical comparison of the five nested, collected sample sizes (100, 50, 20, 10, 5) with the complete 200-specimen dataset.

3. Results

The Mantel Test R-statistics for each dataset are greater than 0.93 ($p < 0.001$) for all subsample proportions with a median sample size greater than 50 individuals (Figure 5-1A). The Pearson's Product Moment correlation coefficients (r-value) between the subsampled and complete datasets' NMDS axis-one scores are greater than 0.86 ($p < 0.001$) for all subsample proportions with a median sample size greater than 50 individuals (Figure 5-1B). The Procrustean Randomization Test (PROTEST) $m^2$-values for each dataset are greater than 0.82 ($p < 0.001$) for all subsample proportions with a median sample size greater than 50 individuals (Figure 5-1C). When the median sample size is less than 50 individuals, the R-statistics, r-values, and $m^2$-values begin to decrease, and then rapidly decrease at median sample sizes smaller than

25 individuals. Between sample sizes of 25 to 50 individuals, the correlation statistics are still significant for most datasets.

There are no patterns or separation among Mantel Test R-statistics, NMDS axis-one score r-values, or PROTEST $m^2$-values among datasets at any sample size based on lithology, time period, initial median sample size, or richness (Tables 5-2 and 5-3; Figure 5-2). There was no significant difference between the mean correlation statistics (R-statistics, r-values, and $m^2$-values) of the two bins of lithology, time periods, complete dataset median sample size, or complete dataset richness at a median sample size of > 50 with two exceptions: NMDS axis-one correlations for the two bins of time periods ($p = 0.02$) and PROTEST for the two bins of complete dataset richness ($p = 0.01$;Table 5-3).

The Rapid Member dataset, which was included to examine the effect of gregariousness, produced results similar to the other 29 datasets when run through the simulation protocol (Figure 5-1). The Mantel R-statistics at a sample size of 50 is R= 0.95, which is greater than that of the lowest of all 30 datasets, R = 0.93. The NMDS axis-one r-value is r = 0.94, which is greater than that of the minimum value out of all 30 datasets, R = 0.86. The PROTEST $m^2$-value is $m^2 = 0.93$, which is greater than that of the minimum value out of all 30 datasets, $m^2 = 0.82$. The nested sampling comparisons of the Rapid Member dataset produce similar results to the 30 datasets run through the simulation protocol with the NMDS axis-one (r = 0.95) and PROTEST ($m^2 = 0.90$) comparisons. However, the Mantel R-statistics at a sample size of 50 was lower than the 30 simulated dataset comparisons (R = 0.84). All of the correlation statistics for both analyses of the Rapid Member dataset subsequently plummet when the sample size is less than 50 individuals (Figure 5-1).

Richness, Pielou's Evenness, and Shannon's H values for subsampled taxon-sample matrices correlate significantly ($p < 0.001$) with their corresponding 100% taxon-sample matrix when the median sample size is greater than 50 individuals. The r-values of the correlation between the richness, Pielou's Evenness, and Shannon's H of the complete dataset and subsample datasets decrease rapidly when the median sample size is less than 25 individuals (Figure 5-3). Between sample sizes of 25 to 50 individuals, the correlation statistics are still significant for most datasets.

4. Discussion

A median sample size of 50 individuals produced the same paleocommunity signal (the information and patterns produced from the taxonomic distribution through and among samples) as all larger sample sizes for all 30 datasets (Figures 5-1 and 5-2). This is evidence that the median required sample size for paleocommunity studies employing multivariate techniques to examine patterns among samples is 50 individuals. It should be noted that this result is specifically geared toward community ecologists attempting to discern paleocommunity differences or gradients using multivariate statistical techniques, and it may not be appropriate

for studies examining trophic interactions or other areas of ecology that require an estimate of the complete diversity of a community.

The Mantel Test comparing dissimilarity matrices produced consistently higher R-statistics than the r-values of the NMDS axis-one scores correlations. This may be due to the conservative nature of the subsampling protocol (i.e., the subsampling protocol was logistically constrained to two-dimensional NMDS ordination and NMDS axis-one correlations alone). The use of a less conservative protocol more appropriate for specific datasets may increase the NMDS correlation statistics closer to that of the Mantel Test.

Lithology, time period, number of taxa, or number of samples did not influence the correlation statistics (Tables 5-2 and 5-3; Figure 5-2). There was no separation based on any of these four parameters when they were mapped onto the plots of the 30 datasets (Figure 5-2). Although two of the 12 t-tests did demonstrate a significant difference between bins of mean correlation statistics, there were no consistent significant differences across two or all three multivariate comparison methods. Thus, the lack of separation based on these four sets of parameters demonstrates that the results for all sample sizes greater than 50 individuals are consistent across all of Phanerozoic time, carbonate and siliciclastic lithologies, all median sample sizes, and all levels of richness. This suggests that none of these variables will affect the required sample size for multivariate paleocommunity research.

4.1 Gregariousness

The method of the subsampling simulation protocol used herein assumed random spatial distribution of taxa throughout each sample. However, gregariousness is prevalent among marine invertebrates (Zuschin et al., 1999; Leighton and Schneider, 2004) and may lead to a non-random distribution of taxa in a sampled stratigraphic horizon (Bennington and Rutherford, 1999; Holland, 2005). In contrast, there is growing evidence that the distribution of taxa within paleocommunities is more random than gregarious (Zuschin et al., 2006) in part due to time averaging (Flessa and Kowalewski, 1994; Kidwell, 2001).

As a test of the effect of gregariousness, both the nested-collected and subsampled results of the Rapid Member dataset demonstrate the same trends (R-statistics, r-values, and $m^2$-values) as one another and as the other 29 datasets (Figure 5-1). , The collected dataset has some of the highest r-values from NMDS axis-one score correlations (Figure 5-1B) and moderate $m^2$-values from the PROTEST analyses (Figure 5-1C). However, the nested-collected dataset has lower Mantel R-statistic values than all of the other datasets (Figure 5-1A). Two possible reasons for this discrepancy between the performance of the nested-collected and the subsampled results when using the Mantel Test are (1) the Mantel subsampled results performed so well that any gregarious real dataset cannot perform as well as the randomized nature of the subsampling protocol. (2) The gregariousness effect may be averaged out, and therefore, not as evident after the data is run through the ordinations. The underlying dataset structure that is used for the

Mantel Test—dissimilarity measures—is more sensitive to the level of gregariousness than are the ordination results (i.e., NMDS axis-one and PROTEST).

Although the gregarious nested-collected dataset did produced lower Mantel Test R-statistics than the subsampled results, the R-statistic at a sample size of 50 is R = 0.84, This is still higher than some of the PROTEST $m^2$-values for median sample sizes > 50. Thus, even when taxa are demonstrated to be gregarious, it may be possible to assess paleocommunity variation using a median sample size of 50 individuals.

Gregarious patches (patchiness) of fossils may lead to inconsistencies in an otherwise consistent environmental or ecological gradient or system. Particularly, if only 50 individuals were collected, there would be no way to assess whether the sample was from a fine-scale patchy distribution or from a coarse-scale heterogeneous population. Therefore, depending on the scale and scope of a study (Forcino et al., 2010b), collecting multiple samples from different positions along each sampled stratigraphic horizon decreases the likelihood of collecting only from fine-scale patches of taxa (Bennington and Rutherford, 1999; Bennington, 2003; Zambito et al., 2008). However, based on the results herein, the required sampling effort at each of these multiple within-stratigraphic-horizon samples is less than the previous recommendations of 200 or more individuals per sample.

4.3 Diversity

The present study demonstrates the same diversity metric patterns and trends among samples in a dataset at all median sample sizes greater than 25 to 50 specimens (Figure 5-3). At a median sample size of 25 to 50 individuals, one may argue diversity metrics (i.e., richness, Pielou's Evenness, and Shannon's H) could be imprecise and lead to erroneous ecological conclusions. However, due to taphonomic processes, the paleocommunity diversity is inherently different than that of the once-living community it represents. Most soft-bodied organisms and many shelled taxa do not end up in the paleocommunity analysis (Cherns and Wright, 2009). However, examining comparable skeletonized taxa, the relative patterns of diversity (i.e., the variation and the relative ranking in richness, Pielou's Evenness, and Shannon's H through samples) are the same as the once-living community (Kidwell, 2002; Olszewski and Kidwell, 2007). Because the patterns are the same, the multivariate-based paleocommunity gradients and patterns can still be obtained with reasonable certainty.

5. Conclusion

Thirty paleocommunity datasets from various time periods, taxonomic groups, scales, and lithologies, containing a wide range of numbers of samples, taxa, and initial median sample sizes were subsampled to five smaller proportional sample sizes. For all 30 datasets, effectively the same multivariate paleocommunity result was found between the complete dataset and proportional subsamples with a median sample size greater than 50 individuals. In addition, the subsampled datasets with a median sample size greater than 50 individuals produce the same

patterns of richness, Pielou's Evenness, and Shannon's H as the corresponding complete dataset. To provide evidence that the gregariousness of taxa does not affect the subsample protocol, one of the 30 datasets subsampled was from the Devonian Rapid Member of Iowa, which is known from previous work to contain gregarious taxonomic communities. The subsampled datasets produced by the Rapid dataset produced the same multivariate result as all sample sizes greater than 50 individuals. This dataset was also collected in a nested fashion, with all smaller sample size included as portions of the larger sample sizes. The multivariate results of the nested collection from the Rapid Member were similar to the 30 subsampled datasets. This demonstrates that in a practical collection for paleocommunity research, a sample size of 50 produces the same multivariate result as all greater sample sizes. Thus, this is evidence that a sample size of 50 individuals is the median required sample size for multivariate paleocommunity research.

6. Acknowledgements

This paper was supported by the Paleontological Society G. Arthur Cooper Award. I would like to thank Lindsey Leighton for his guidance and feedback throughout the span of this project, Emily Stafford, Chris Schneider, and Brian Chatterton for their manuscript reviews, Kristina Barclay and Simen Linge-Johnsen for their assistance collecting and identifying specimens from the Rapid Member samples, and J.C. Cahill for additional feedback. A special thank you to Tom Olszewski and Andy Bush for their reviews and input leading to additional analyses that added to the strength of this research.

7. Bibliography

Alroy J., Aberhan, D.J., Bottjer, D., Foote, F.T., Fürsich, P.J., Harries, A.J.W., Hendy et al., 2008. Phanerozoic trends in the global diversity of marine invertebrates. Science. 321, 97–100.

Bennington, J.B., 2003. Transcending patchiness in comparative analysis of paleocommunities: a test case from the Upper Cretaceous of New Jersey. PALAIOS. 18, 22–33.

Bennington, J.B. and Rutherford, S.D., 1999. Precision and reliability in paleocommunity comparisons based on cluster-confidence intervals: how to get more statistical bang for your statistical buck. PALAIOS. 14, 506–515.

Bonelli Jr., J.R., Brett, C.E., Miller, A.I. and Bennington, J.B., 2006. Testing for faunal stability across a regional biotic transition: quantifying stasis and variation among recurring coral-rich biofacies in the Middle Devonian Appalachian Basin. Paleobiology. 32, 20–37.

Bonelli Jr., J.R. and Patzkowsky, M.E., 2008. How are global patterns of faunal turnover expressed at regional scales? Evidence from the Upper Mississippian (Chesterian Series), Illinois Basin, USA. PALAIOS. 23, 760–772.

Bulinski, K.V., 2007. Analysis of sample-level properties along a paleoenvironmental gradient: The behavior of evenness as a function of sample size. Palaeogeography, Palaeoclimatology,

Palaeoecology. 253, 490–508.

Bunker, B.J. and Witzke, B.J., 1992. An upper Middle through Upper Devonian lithostratigraphic and conodont biostratigraphic framework of the Midcontinent Carbonate Shelf area, Iowa, in: Day, J., Bunker, B.J. (Eds.), The Stratigraphy, Paleontology, Depositional and Diagenetic History of the Middle-Upper Devonian Cedar Valley Group of Central and Eastern Iowa: Iowa City, Iowa, Iowa Department of Natural Resources, pp. 3–26.

Bush, A.M. and Brame, R.I., 2010. Multiple paleoecological controls on the composition of marine fossil assemblages from the Frasnian (Late Devonian) of Virginia, with a comparison of ordination methods. Paleobiology. 36, 573–591.

Caron J.B. and Jackson, D.A., 2008. Paleoecology of the Greater Phyllopod Bed community, Burgess Shale. Palaeogeography, Palaeoclimatology, Palaeoecology. 258, 222–256.

Chang, Y.M., 1967. Accuracy of fossil percentage estimation. Journal of Paleontology. 41, 500–502.

Cherns, L. and Wright, V.P., 2009. Quantifying the impacts of early diagenetic aragonite dissolution on the fossil record. PALAIOS. 24, 756–771.

Clapham, M.E. and Bottjer, D.J., 2006. Permian marine paleoecology and its implications for large-scale decoupling of brachiopod and bivalve abundance and diversity during the Lopingian (Late Permian). Palaeogeography, Palaeoclimatology, Palaeoecology. 249, 283–301.

Clapham, M.E., Bottjer, D.J., Powers, C.M., Bonuso, N., Fraiser, M.L, Marenco, P.J., Dornbos, S.Q. and Pruss, S.B., 2006. Assessing the ecological dominance of Phanerozoic marine invertebrates. PALAIOS. 21, 431–4411.

Clarke, K.R. and Ainsworth, M., 1993. A method and linking multivariate community structure to environmental variables. Marine Ecology. 92, 205–219.

Fall, L.M. and Olszewski, T.D., 2010. Environmental disruptions influence taxonomic composition of brachiopod paleocommunities in the Middle Permian Bell Canyon Formation (Delaware Basin, West Texas). PALAIOS. 25, 247–259.

Fatela, F. and Taborda, R., 2002. Confidence limits of species proportions in microfossil assemblages. Marine Micropaleontology. 45, 169–174.

Finnegan, S. and Droser, M.L., 2005. Relative and absolute abundance of trilobites and rhynchonelliform brachiopods across the Lower/Middle Ordovician boundary, eastern Basin and Range. Paleobiology. 21, 480–502.

Flessa, K.W, and Kowalewski, M., 1994. Shell survival and time averaging in nearshore and shelf environments: estimates from the radiocarbon literature. Lethaia. 27, 153–165.

Forcino, F.L., Richards, E.J. Chojnacki, N., Stafford, E.S. and Leighton, L.R., 2010b. Spatial variation of late Pennsylvanian paleocommunities: ecological and sampling implications. Geological Society of America Abstracts with Programs. 42, 27.

Forcino, F.L., Stafford, E.S., Warner, J.J., Webb, A.E., Leighton, L.R., Schneider, C.L., Michlin, T.S., Palazzolo, L.M., Morrow, J.R. and Schellenberg. S.A., 2010a. Effects of data categorizations on paleocommunity analysis: a case study from the Pennsylvanian Finis Shale of Texas. PALAIOS. 25, 144–157.

Hayek, L.C. and Buzas, M.A., 1997. Surveying Natural Populations. Columbia University Press, New York.

Heck, Jr., K.L., van Belle, G. and Simberloff, S., 1975. Explicit calculation of the rarefaction diversity measurement and the determination of sufficient sample size. Ecology. 56, 1459–1461.

Heim, N.A., 2009. Stability of regional brachiopod diversity structure across the Mississippian/Pennsylvanian boundary. Paleobiology. 35, 393–412.

Holland, S.M., 2005. The signature of patches and gradients in ecological ordinations. PALAIOS. 20, 573-580.

Jackson, D.A., 1995. PROTEST: A Procrustean randomization test of community environment concordance. Écoscience. 2, 297–303.

Kidwell, S.M., 2001. Preservation of species abundance in marine death assemblages. Science. 294, 1091–1094.

Kidwell, S.M., 2002. Time-averaged molluscan death assemblages; palimpsests of richness, snapshots of abundance. Geology. 30, 803–806.

Kowalewski, M., Gurs, K., Nebelsick, J.H., Oschmann, W., Piller, W.E. and Hoffmeister, A.P., 2002. Multivariate hierarchical analyses of Miocene mollusk assemblages of Europe: Paleoceanographic, paleoecological, and biostratigraphic implications. GSA Bulletin. 114, 239–256.

Kruskal, J.B., 1964. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychomerika. 29, 1–27.

Leighton, L.R. and Schneider, C.L., 2004. Neighbor Proximity Analysis, A Technique for Assessing Spatial Patterns in The Fossil Record. PALAIOS. 19, 396–407.

Magurran, A.E., 2004. Measuring Biological Diversity. Blackwell Publishing, Oxford, England.

Mantel, N., 1967. The detection of disease clustering and a generalized regression approach. Cancer Research. 27, 209–220.

McCune, B. and Grace, J.B., 2002. Analysis of Ecological Communities. MjM Software Design, Gleneden Beach, Oregon.

Miller, A.I. and Foote, M., 1996. Calibrating the Ordovician radiation of marine life: implications for Phanerozoic diversity trends. Paleobiology. 22, 304–309.

Minchin, P.R., 1986. An evaluation of the relative robustness of techniques for ecological ordination. Plant Ecology. 69, 89–107.

Oksanen, J., Blanchet, F.G., Kindt, R., Legendre, P., O'Hara, R.B., Simpson, G.L., Solymos, P., Stevens, M.H.H. and Wagner, H., 2010. vegan: Community Ecology Package. R package version 1.17-3, http://CRAN.R-project.org/package=vegan.

Olszewski, T.D. and Kidwell, S.E., 2007. The preservational fidelity of evenness in molluscan death assemblages. Paleobiology. 33, 1–23.

Olszewski, T.D. and Patzkowsky, M.E., 2001. Measuring recurrence of marine biotic gradients: a case study from the Pennsylvanian-Permian Midcontinent. PALAIOS. 16, 444–460.

Patterson, R.T. and Fishbein, T., 1989. Re-examination of the statistical methods used to determine the number of point counts needed for micropaleontological quantitative research. Journal of Paleontology. 63, 245–248.

Peres-Neto, P.R. and Jackson, D.A., 2001. How well do multivariate datasets match? The advantages of a Procrustean superimposition approach over the Mantel Test. Oecologia. 129, 169–178.

Phleger, , F.B., 1960. Ecology and Distribution of Recent Forminifera. Johns Hopkins Press, Baltimore, MD.

R Development Core Team, 2010. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Raup, D.M., 1975. Taxonomic diversity estimation using rarefactions. Paleobiology. 1, 333–342.

Redman, C.L., Leighton, L.R., Schellenberg, S.A., Gale, C.N., Neilsen, J.L., Dressler, D.L., and Klinger, M.K., 2007. Influence of spatiotemporal scale on the interpretation of paleocommunity structure: lateral variation in the Imperial Formation of California. PALAIOS. 22, 630-641.

Tomasovych, A., 2006. Brachiopod and bivalve ecology in the Late Triassic (Alps, Austria): Onshore-offshore replacements caused by variations in sediment and nutrient supply. PALIOS. 21, 344–368.

Wagner, P.J., 2006. Abundance distribution imply elevated complexity of Post-Paleozoic marine ecosystems. Science. 314, 1289–1292.

Zambito IV, J.J., Mitchell, C.E., and Sheets H.D., 2008. A comparison of sampling and statistical techniques fro analyzing bulk-sampled biofacies composition. PALAIOS. 23, 313–321.

Zuschin, M., Harzhauser, M., and Sauermoser, K., 2006. Patchiness of local species richness and its implication for large-scale diversity patterns: an example from the middle Miocene of the Paratethys. Lethaia. 39, 65–78.

Zuschin, M., Stachowitsch, M., Pervesler, P., and Kollmann, H., 1999. Structural features and taphonomic pathways of a high-biomass epifauna in the northern Gulf of Trieste, Adriatic

Seas. Lethaia. 32, 299–317.

**Table 5-1.** Listing of the 30 datasets to which the subsampling code was applied. Provided is the reference citation, PBDB reference number when applicable, time period, primary lithology, number of genera, number of samples, and median sample size of the complete dataset.

| Reference | PBDB ref # | Time Periods | Primary Lithology | # of Genera | # of Samples | 100% Median Sample Size |
|---|---|---|---|---|---|---|
| Forcino et al., 2010 | N/A | Pennsylvanian | Shale | 51 | 13 | 323 |
| Forcino et al., in review | N/A | Pennsylvanian | Shale | 81 | 33 | 145 |
| Johnson and Klapper, 1990 | N/A | Devonian | Sandstone | 167 | 24 | 2441 |
| Forcino et al., in preparation | N/A | Devonian | Limestone | 39 | 15 | 200 |
| Ausich, personal data. | N/A | Mississippian | Shale | 101 | 15 | 366 |
| Popov et al., 2002 | 26964 | Ordovician | Sandstone | 72 | 41 | 21 |
| Heim, 2009 | 26838 | Miss-Penn | Limestone | 45 | 61 | 28 |
| Bulinski, 2007 | 24916 | Ordovician | Limestone | 24 | 140 | 26.5 |
| Holland and Patzkowsky, 2004 | 9838 | Ordovician | Limestone | 37 | 93 | 45 |
| Budd et al., 1999 | 24129 | Neogene | Limestone | 31 | 112 | 15 |
| Fürsich and Wendt, 1977 | 23366 | Triassic | Limestone | 159 | 51 | 61 |
| Peters, 2003 | 13544 | Cam-Ord | Limestone | 120 | 30 | 166 |
| Schneider, 2003 | 11551 | Pennsylvanian | Shale | 34 | 72 | 39.5 |
| Boomer, et al., 1998 | 8881 | Jurassic | Limestone | 24 | 31 | 161 |
| Gahr. 2002 | 7068 | Jurassic | Limestone | 100 | 144 | 67 |
| Holzapfel, 1998 | 6119 | Jurassic | Limestone | 107 | 96 | 65.5 |
| Patzkowsky, 1995 | 401 | Ordovician | Limestone | 25 | 44 | 47 |
| Williams, et al., 1981 | 162 | Ordovician | Sandstone | 29 | 76 | 78 |
| Scott, 1970 | 52 | Cretaceous | Sandstone | 60 | 75 | 27 |
| Schubert and Bottjer, 1995 | 8833 | Triassic | Limestone | 31 | 34 | 69 |
| Watkins, 1994 | 4161 | Silurian | Limestone | 32 | 19 | 99 |
| Waterhouse, 1987 | 11597 | Permian | Sandstone | 71 | 74 | 46.5 |
| Dominici and Kowalke. 2007 | 25442 | Paleogene | Shale | 72 | 99 | 103 |
| Clapham, 2006 | 13431 | Permian | Limestone | 133 | 15 | 43 |
| Ahmad, 1998 | 9529 | Jurassic | Limestone | 50 | 52 | 575 |
| Crame. 1981 | 18189 | Neogene | Limestone | 28 | 38 | 9 |
| Hogler, 1992 | 8769 | Triassic | Limestone | 38 | 85 | 37 |
| Etter, 1990 | 9638 | Jurassic | Shale | 36 | 124 | 182 |
| Tomasovych, 2006 | 17562 | Triassic | Limestone | 21 | 20 | 46.5 |
| Stilwell et al., 2004 | 19080 | Paleogene | Sandstone | 38 | 38 | 10 |

**Table 5-2.** Mean, minimum (min), and maximum (max) correlations statistics (R-statistics for Mantel Test, r-values for NMDS axis-one correlations, and $m^2$-values for PROTEST) for four groupings: lithology (carbonate versus siliciclastic), time, complete dataset median sample size, and complete dataset richness. The shading separates the pairs of values that are compared.

| | Mantel | | | NMDS Axis-one | | | PROTEST | | |
|---|---|---|---|---|---|---|---|---|---|
| | Mean | Min | Max | Mean | Min | Max | Mean | Min | Max |
| Carbonate | 0.98 | 0.95 | 0.99 | 0.95 | 0.88 | 0.98 | 0.89 | 0.82 | 0.96 |
| Siliciclastic | 0.99 | 0.96 | 0.99 | 0.97 | 0.92 | 0.97 | 0.93 | 0.82 | 0.99 |
| Paleo | 0.98 | 0.95 | 0.99 | 0.95 | 088 | 0.99 | 0.92 | 0.82 | 0.99 |
| Post-Paleo | 0.99 | 0.98 | 0.99 | 0.98 | 0.96 | 0.99 | 0.89 | 0.83 | 0.99 |
| Small Sam | 0.97 | 0.96 | 0.99 | 0.91 | 0.88 | 0.95 | 0.88 | 0.82 | 0.93 |
| Large Sam | 0.98 | 0.95 | 0.99 | 0.96 | 0.88 | 0.99 | 0.92 | 0.83 | 0.99 |
| Low S | 0.98 | 0.96 | 0.99 | 0.97 | 0.94 | 0.99 | 0.96 | 0.93 | 0.99 |
| High S | 0.99 | 0.96 | 0.99 | 0.95 | 0.88 | 0.99 | 0.90 | 0.82 | 0.99 |

**Table 5-3.** T-test p-values for comparisons of the two groups of correlation statistics from median sample sizes greater than 50 individuals for each of the four sets of parameters.

| | Lithology | Time | Sample Size | Richness |
|---|---|---|---|---|
| Mantel | p = 0.51 | p = 0.32 | p = 0.52 | p = 0.65 |
| NMDS axis 1 | p = 0.23 | p = 0.02 | p = 0.44 | p = 0.35 |
| PROTEST | p = 0.10 | p = 0.41 | p = 0.58 | p = 0.01 |

**Figure 5-1.** The solid black line is the median sample size (number of individual fossil specimens per sample) for each proportion of each dataset versus [A] the corresponding mean Mantel R-statistic, [B] the corresponding mean Pearson's Product Moment correlation coefficient (r-value) for NMDS axis-one scores, and [C] the corresponding mean PROTEST $m^2$-value. The dashed grey lines are the 95% confidence extremes (plus and minus two standard deviations). The vertical grey bar in the background of the plots marks the location of median samples sizes 25 to 50. The arrow in each plot is pointing to the Iowa Rapid Member dataset (circle points) that was field sampled at various sizes.

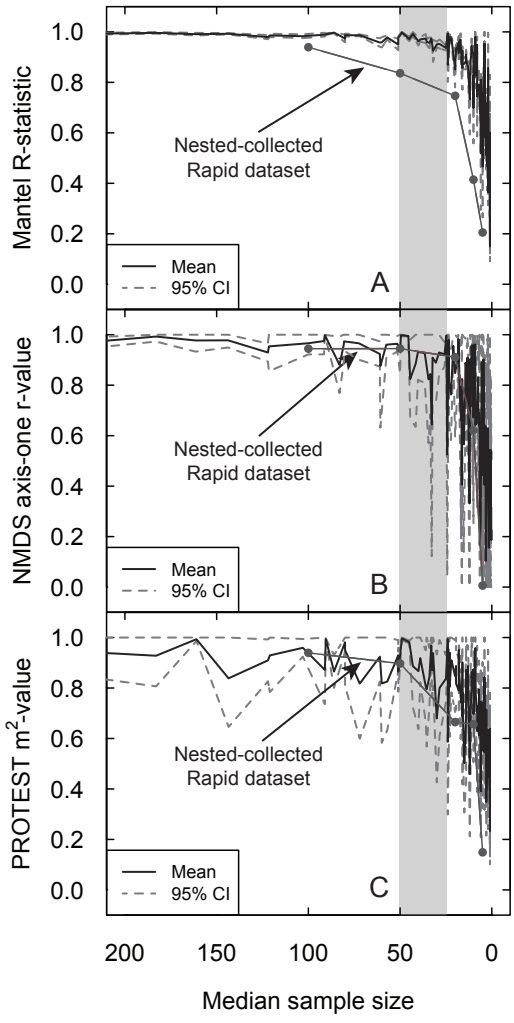


**Figure 5-2.** Median sample size (number of individual fossil specimens per sample) for each proportion of each dataset versus the corresponding mean Pearson's Product Moment correlation coefficient (r-value) for NMDS axis-one scores for each of the 30 datasets plotted individually. [A] Each line is shaded based on lithology, with the black line representing siliciclastics and the grey line representing carbonates. [B] Each line is shaded based on time with the black line representing the Cambrian through Permian (Paleozoic) and the gray line representing Triassic through Neogene (Post-Paleozoic). [C] Each line is shaded based on the median sample size of the complete dataset with the black line representing the 15 datasets with the larger median sample and the grey line representing the 15 datasets with the smaller median sample size. [D] . Each line is shaded based on the richness of the complete dataset with the black line representing the 15 datasets with the higher richness and the grey line representing the 15 datasets with the lower richness. For all four plots, the vertical grey bar in the background of the plots marks the location of median samples sizes 25 to 50.

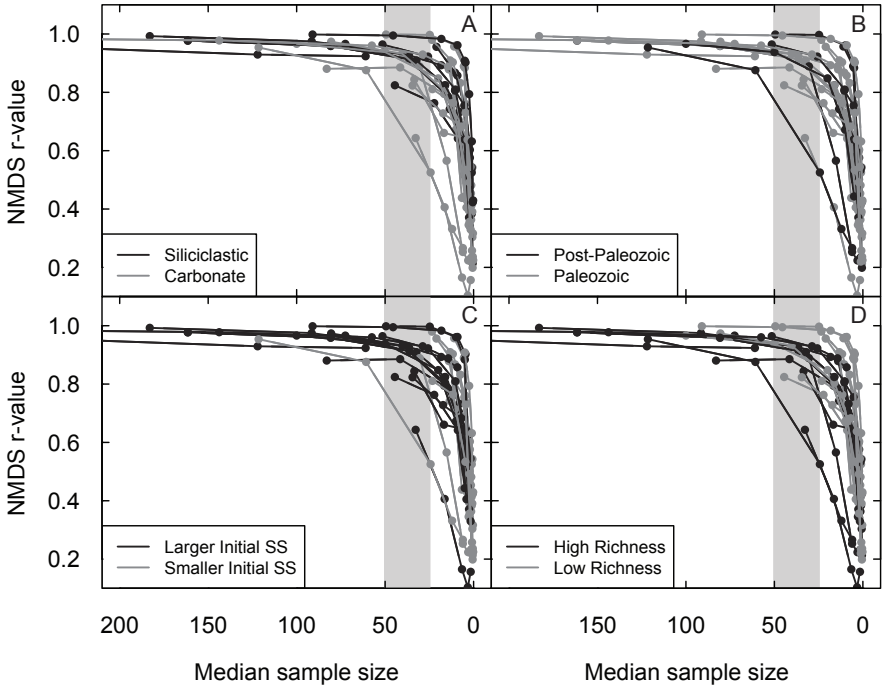
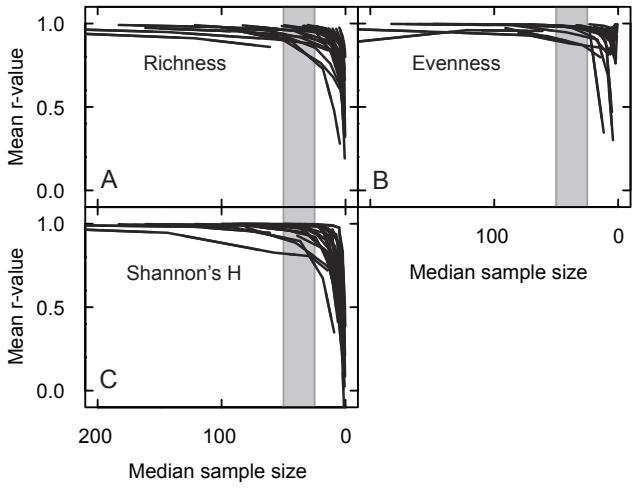


**Figure 5-3.** Median sample size for each proportion of each dataset versus corresponding r-values between the 100% matrix for all five proportions and [A] mean richness (R), [B] mean Buzas and Gibson's evenness (E), and [C] mean Shannon's Enthalpy (H) of each of the 1000 subset taxon-sample matrices. The vertical grey bar in the background of the plots marks the location of median samples sizes 25 to 50.



# Chapter 6. Sample size requirements for multivariate abundance-based community research: meta-analysis and model-based approach[**]

## 1. Introduction

Many ecological studies use multivariate techniques (e.g., ordination, analysis of similarity) to assess patterns and gradients of taxonomic variation (Warwick and Clarke 1993, Legendre and Legendre 1998, McCune and Grace 2002, Forcino et al. 2010). To obtain the most accurate and statistically viable information from a taxonomic series of samples, ecologists must ensure they are collecting a representative sample of the community or series of samples in question (Hayek and Buzas 1997, Jalonen 1998, Bennington and Rutherford 1999, Cao et al. 2002, Schloss 2008). An important component of the statistical viability of an ecological sample is the number of individuals per sample—sample size. A sample size, for our purposes, is the total number of individual specimens comprising one row of data in a taxon-sample matrix used for multivariate community analysis; in many studies, the sample is presumed to represent a community. Without a sample size large enough to be representative of the entire community, erroneous ecological conclusions can be reached. However, it is important not to over-sample, because an increasingly larger sample size will eventually result in diminishing returns in terms of improving any pattern or gradient revealed by the data; resources and effort would be wasted. Here, I compare the multivariate statistical results of subsampled datasets to corresponding complete datasets with the goals of determining (1) if smaller sample sizes produce the same results as larger sample sizes and (2) the types of datasets (the structure and parameters) that require larger sample sizes. Although smaller sample sizes may result in a probability of not collecting rare taxa, one of the strengths of using multidimensional analyses of communities is that they may still retain relative taxonomic information among communities; even with smaller samples, individual communities may still plot in the same relative positions in ordination space, allowing for identification of the same community trends and patterns.

Previous research has determined appropriate sample sizes for ecological research by examining the probability of acquiring species that comprise some proportion of a sample with 95% confidence (Phleger 1960, Chang 1967, Patterson and Fishbein 1989, Fatela and Taborda 2002, Choa et al. 2005, Schloss and Handelman 2007). In other words, how likely is a sample to contain taxon Z if a sample size of 300 individuals is collected? Examining probability in this manner does not take into account relationships between sampling units or taxa (e.g., the interaction of taxa within a community). Others have compared multivariate results of plots or samples collected at different sample sizes to determine which sample size captures the most information (Goff and Mitchell 1975, Cao et al. 2002, Ranjard 2003, Kang and Mills 2006,

[**] *A version of this chapter has been submitted for publication. Forcino, F.L., Twerdy, P., and Leighton, L.R. (in review) Sample size requirements for multivariate abundance-based community research: meta-analysis and model-based approach. Methods in Ecology and Evolution.*

Zdenka and Milan 2006). Although these case studies take a more practical approach than the probability estimates, too few exist to lead to the creation of a standard sampling protocol that encompasses multiple different types of ecological research.

Forcino (2012) conducted a meta-analysis of 30 fossil community datasets from a wide range of time periods, taxonomic groups, spatial and temporal scales, lithologies, number of samples (13–124), and number of taxa (21–167), and median sample sizes (9–2441). He randomly subsampled without replacement each sample within each dataset and compared the multivariate result (ordination or other multivariate technique) of the subsample to that of its corresponding complete dataset. This procedure was repeated 1000 times to five proportional sizes (50%, 25%, 10%, 5%, 2.5%) for each dataset. The results demonstrated that a median sample size of 50 individuals per collected sample (or per fossil community) produced the same multivariate results as those of all larger sample sizes. This method of comparing the results of multivariate techniques applied to community datasets at different sample sizes has the greatest practical utility because the use of multivariate techniques is so prevalent in the literature.

Here, using similar methods as Forcino (2012), I examined 396 datasets and statistically compared the multivariate results of subsamples to the multivariate results of the corresponding complete datasets using three different types of ecological data: 44 real datasets (from a range of taxonomic groups, geographic locations, spatial scales, and environments), 220 simulated-created datasets, and 132 selected-created datasets (see Methods). The two goals of this research are to (1) test further the hypothesis of Forcino (2012) that smaller samples sizes (as few as 50 individuals per sample) produce the same multivariate results as larger sample sizes and (2) determine the dataset parameters (e.g., number of taxa, number of samples) that lead to a lack of correlation between subsampled and complete datasets, and therefore, require larger sample sizes. Additional evidence to support goal (1) will provide community ecologists with a more accurate representative sample size for multivariate research, and lead to a better use of resources. Accomplishing goal (2) will inform ecologists when greater resources are necessary to obtain a statistically viable sample of a community.

## 2. Methods

### 2.1 Real datasets

Forty-four real datasets were acquired from the ecological literature (Table 6-1). These datasets consist of various numbers of taxa (3 to 421), numbers of samples (4 to 445), were from different taxonomic groups, different geographic locations, and different environments (Table 6-1). Real datasets were used in order to determine if datasets used in previous studies would produce the same multivariate community results if fewer individuals were sampled. Six of the 44 datasets selected were from publications in *Ecology* from the years 2008 to 2011. In addition, 18 datasets were gathered from one meta-analysis in *Ecology* (Ulrich and Gotelli, 2010). These

18 datasets are cited from each original study, and not the meta-analysis. Twenty studies from other journals were also used to increase the number of datasets (Table 6-1). By selecting from recent issues of *Ecology* I was able to gain a representation of various types of communities from various taxonomic groups and authors. I randomly selected and subsampled the first 44 datasets that I could find that had median sample sizes of at least 20 individuals.

2.2 Simulated-created dataset construction

I created 220 simulated-created datasets that were similar to real datasets. However, there is an additional random component to the simulated-created datasets (see below) along with some control on the parameters compared to the real dataset where I controlled none of the parameters or values. Furthermore, I wanted to examine the values within the sets of parameters (e.g., sample sizes larger than 1000 individuals per sample) that did not exist within the 44 real datasets. Similar to the real datasets, these datasets contained a range of numbers of taxa and samples (Table 6-2).

The total number of 220 simulated-created datasets and 44 real datasets was arbitrary. I could have added additional datasets if I thought it would reveal additional insight into sample size requirements. However, I stopped at 220 and 44 in order to balance the amount of time it would take to gather and subsample the datasets versus the amount of information that would be provided by each dataset.

Datasets were simulated by first randomly creating a normal distribution of abundances for each taxon across a hypothetical gradient (Figure 6-1). Each normal distribution for each taxon was created based on a randomly chosen mean, randomly chosen standard deviation, and randomly chosen maximum frequency. The resulting distribution represents the range along an ecological or environmental gradient within which each simulated taxon is located. Each simulated taxon has a peak value (maximum frequency) where its abundance is at its maximum (located at the mean of the normal distribution), and areas where the taxon is less likely to be found (the tails of the normal distribution). For example, if the ecological gradient represents water depth, the randomly selected mean of the normal distribution for Taxon A represents the optimal depth at which Taxon A lives, and therefore, maximum peak of abundance. The tails of the normal distribution represent the most extreme conditions (shallowest and deepest depths) in which Taxon A lives, with abundances declining from the peak to each tail.

Each normal distribution was randomly generated then placed along the simulated-environmental gradient for each taxon in that particular simulation; subsequently, sample locations were randomly selected along the environmental gradient (Figure 6-1), simulating random sampling of a gradient in the field under circumstances where continuous or interval sampling is not possible. For example, if the gradient was 100 units long (the unit length is an arbitrary value representing the complete gradient length), a unique number from 1 to 100 was randomly selected for each sample, which represents the sampling locations along the created

environmental gradient. At each sampled point along the gradient, all of the taxonomic distributions that cross that point are included in that sample. Taxon abundances equal the height of the curve of each taxon's normal distribution at that point along the gradient (Figure 6-1).

2.3 Selected-created dataset construction

Lastly, 132 selected-created datasets were constructed with the intention of deliberately generating more extreme differences among complete and subsampled datasets. In each dataset, the authors selected the abundances (as opposed to simulated the abundances) for each taxon in each sample, and parameters were varied systematically to test for differences in the number of samples (i.e., the number of collected samples in a dataset that would equate to one row in a taxon by sample matrix), numbers of taxa, and evenness. These three variables were selected because they commonly vary from study to study, they are the some of the most basic ecological measures (i.e., number of taxa and evenness), and they are easy to control in analyses.

This selected, systematic creation of datasets often led to datasets with rank abundance distribution and absolute abundances that are rare in the literature, but which might facilitate identification of the conditions under which larger sample sizes would be necessary to capture a representative dataset.

Each sample of each selected-created dataset contained 200 individuals. A sample size of 200 was chosen so that the five subsample proportion sizes (i.e., 100, 50, 20, 10, 5) could be produced that represent a range of sample sizes. These sample sizes of 5 to 200 are commonly used in ecological research.

I systematically varied the number of samples, richness, and evenness when creating the 132 selected-created datasets in order to examine these three parameters as possible influences on the required sample size for community research. By varying the number of samples in each dataset, I defined 81 datasets containing 5 samples and 51 datasets containing 10 samples. Five and 10 were chosen in order to keep the datasets simple while still being able to evaluate whether the number of samples had an effect on the required sample size. The 5- or 10-sample datasets represent real datasets of 5 or 10 different communities, respectively. In some community studies, some samples will have a similar enough taxonomic makeup to be indistinguishable in a multivariate analyses; these samples would be considered the same community. Because all 5 or 10 samples in the 132 constructed datasets have a different taxonomic makeup and are designed to be distinguishable in multivariate analyses, each of the 5 or 10 samples represents a distinct community.

The second parameter that was varied among the 132 datasets was taxonomic richness within each sample. Within the datasets of 5 samples, 27 have a richness of 10, 27 have a richness of 20, and 27 have a richness of 50. Within the datasets of 10 samples, 24 have a richness of 10 and 27 have a richness of 20.

The third parameter varied among the 132 datasets was the evenness. There were three evenness categories: (1) the 44 low evenness datasets, with individual-sample evenness values ranging from 0.14 to 0.98 and a mean evenness of 0.58 (Pielou's J Evenness; Magurran 1998); (2) the 44 high evenness datasets with individual-sample evenness values from 0.18 to 1.0, with a mean of 0.79; and (3) the 44 mixed evenness datasets, with each dataset containing individual samples of both high and low evenness, with individual-sample evenness values from 0.13 to 1.0, with a mean of 0.69. Each sample in each dataset had its evenness varied by starting with one taxon with the greatest abundance while the remaining taxa all had an abundance of 1. For example, a dataset with 5 samples and 10 taxa would have one sample with one taxon having a sample of 191 while the remaining taxa all had an abundance of 1. This example would have been an extreme case of low evenness. Each abundance distribution in each sample of each dataset systematically varied until a dataset contained one sample with every taxon having the same abundance. This example would have been an extreme case of high evenness.

A set of parameters incorporates the three controlled parameters (samples, richness, and evenness). An example set of parameters is 5 samples, a richness of 10, and low evenness. The two sample parameters (5 and 10), the three richness parameters (10, 20, and 50), and the three evenness parameters (high, low, and mixed) combined to create 16 unique sets of parameters. These 16 sets lead to 16 primary datasets. Each primary dataset was set up as a typical community ecology dataset, with a series of samples each containing a number of taxa with varying abundances. The 16 primary datasets do not contain any taxonomic abundances of zero in any of the samples. No datasets with zero were included in order to keep the dataset variants relatively simple; once an abundance of zero was added for any taxon in any sample, it would have had to be systematically varied throughout. This would have led to an exponentially greater number of datasets requiring subsampling, and therefore, an exponentially greater amount of time added to the study.

Various dataset structures were created in order to examine the effect that changing a particular dataset with respect to a certain set of parameters has on the required sample size. A dataset structure*,* for the purposes of this study, refers to the variation of the 16 sets of parameters created by varying taxonomic placement within a sample. Each of these structures creates a variant of the 16 primary datasets, to produce 8 to 9 different variations of the 16 primary datasets equaling 132 total datasets. The first dataset structure was simply the primary dataset of each of the 16 sets of parameters. Each of the remaining 7 or 8 dataset structures manipulated the primary dataset structure in some manner.

The second structure was created by varying the presence or absence of the single most abundant taxon within each sample of the first structure (primary dataset). For example, within the primary dataset that consists of 5 samples, a richness of 10, and low evenness, the same taxon is the most abundant taxon in all five samples. To create the second structure, an additional four

taxa are added to the dataset of the first structure. For each of the four new taxa, an abundance of zero is added to four samples, while in the fifth sample, the value of the highest taxonomic abundance of that sample is added. The result is four new samples, each with one new taxon replacing the original most abundant taxon, and one unchanged sample (with an abundance of zero for all four new taxa).

For the next two structures (the third and fourth), the remaining taxa in each sample—all but the most abundant taxon—were divided into thirds. The first third contained the moderately abundant taxa, and the second two-thirds contained the rare taxa.

For the third structure, the moderately abundant taxa were varied in the same manner as the most abundant taxon in the second structure. The difference between the second and third structures is all of the taxa varied in concert, as opposed to one taxon at a time; each group of three taxonomic abundances was either present or absent in the remaining samples. The most abundant taxon and the rare taxa remained constant through all the samples in the datasets with this structure.

For the fourth structure, the rare taxa were varied in the same manner as the most abundant taxon in the second structure and the moderately abundant taxa of the third structure. The most abundant taxon and moderately abundant taxa remained constant through all the samples in the datasets with this structure.

The fifth and sixth structures changed the position of the most abundant taxon through the five or 10 samples. The most abundant taxon would change from taxon A to taxon B from sample 1 to sample 2 and from taxon A to taxon C from sample 1 to sample 3. The reason this manipulation was occasionally done to two different structures (fifth and sixth or just fifth) was because datasets with only 5 samples and 10 taxa could only have the most abundant taxon vary through 5 different locations. The sixth structure placed the most abundant taxa in the position once occupied by taxa with lower abundances. The sixth structure is the one structure that was not possible to create for all the different parameters. When there were 10 samples and 10 taxa, switching the most abundant taxa would take up all of the taxa with just one structure.

The final three structures were all randomized dataset constructions. The position of each taxon's abundance from the primary dataset (the first structure) was randomly placed in any one of the taxonomic positions. The difference between the three structures was the number of zeros available to add and the total number of taxa. Structure seven contained no zeros with the same number of taxa as the first six structures. Structure eight added an additional possible 5, 10, or 15 zeros for the 10, 20, and 50 taxa datasets, respectively, with the same number of taxa as the first six structures. The ninth and final structure added an additional possible 5, 10, or 15 zeros for the 10, 20, and 50 taxa datasets, respectively, and also added that number of taxa to the dataset so no abundance was unused.

2.4 Subsampling protocol

Using R 2.14 (R Development Core Team 2012), each sample within each taxon-sample matrix (simulated-created, selected-created, and real) was randomly subsampled without replacement to five proportional sizes: 50%, 25%, 10%, 5%, and 2.5% of the total number of individuals in the original sample. For each of the subsampled proportions of each taxon-sample matrix, 1000 subsampled matrices were constructed for a total of 5000 subsampled matrices for each dataset. Each of the 5000 subsampled matrices was statistically compared to the original 100% taxon-sample matrix using two multivariate statistical methods.

(1) Using the vegan package in R 2.14 (Oksanen et al. 2012), Mantel Tests of correlation were performed between the Bray-Curtis dissimilarity matrices (measures of the differences between each object in a taxon-sample matrix) of subsamples and corresponding complete datasets. The Mantel Test tests the similarity of two matrices of dissimilarity indices by permuting each of the elements in the dissimilarity matrix 999 times, to derive a distribution of correlation values (Mantel 1967; Fall and Olszewski 2010). The resulting R-statistic is analogous to the Pearson's Product Moment Correlation Coefficient (r); with increasingly similar data matrices, the Mantel R-statistic will approach 1.

(2) For each of the datasets and subsamples, non-metric multidimensional scaling (NMDS) ordinations of the samples were performed using the Bray-Curtis dissimilarity index (Clarke and Ainsworth 1993; Legendre and Legendre 1998; McCune and Grace 2002; Bush and Brame 2010). All ordinations were run examining the taxonomic distributions among samples with two dimensions with "autotransform = alse" in the vegan package in R, specifically using the function "metaMDS()".

Procrustean Randomization Tests (PROTEST) were performed comparing procrustes transformed ordinations of the subsampled and corresponding complete datasets (Jackson 1995; Peres-Neto and Jackson 2001). NMDS does not always assign the maximum explanation of variation in the ordination space to the first axis. Moreover, two different ordinations might not appear to be similar at first because they are close reflections to each other in ordination space. To address these possibilities, the first step in PROTEST is to perform a Procrustes transformation, which minimizes the sum-of-squares deviations between the two ordination results through translation, reflection, rotation, and dilation. Thus, the two ordination results are reoriented such that they are aligned as closely as possible in ordination space, which permits a more accurate assessment of similarity. The residuals between the two ordinations post-transformation are calculated and produce the $m^2$-value. The $m^2$-value is similar to the r-value resulting from a Pearson's Product Moment Correlation; the closer $m^2$ is to 1, the more similar the two ordinations. Subsequent to the Procrustes transformation, PROTEST randomly permutes the ordination scores for all samples for 999 iterations, and the $m^2$-value is calculated for each iteration; a realized p-value, indicating the significance of the $m^2$-value, is then calculated by

determining the percentage of iterations in which the $m^2$-values from the randomized iterations is greater than the $m^2$-value for the actual dataset.

2.5 Qualitative Assessment

Goodness-of-fit statistics (i.e., Mantel R-statistics and PROTEST $m^2$-values) quantify the similarity between multivariate datasets, but do not indicate whether two ordinations would lead a researcher to draw the same ecological conclusions between both ordination results. While a goodness-of-fit statistic value of 0.8 clearly indicates greater similarity than 0.6, it is not immediately evident just from the number whether a result of 0.8 indicates sufficient similarity that the result would lead a researcher to conclude the two ordinations are "the same". Two ordinations' plots may overall be mathematically similar but important details may differ. For example, do the exact same samples form a recognizably discrete grouping in both ordinations? Are trends within the ordination plot non-linear? Evaluating the patterns visually is an important check, analogous to a visual inspection of any data plot.

Similarly, the p-values derived from the Mantel Test and PROTEST do not necessarily indicate statistical significance in the same manner as a p-value for a classical statistical test such as a Pearson Product Moment Correlation. Specifically, a $p < 0.05$ does not necessarily indicate significant similarity of a subsampled and corresponding complete dataset, in the sense that the two datasets would be interpreted similarly. The p-values in both of these tests are calculated by determining if the observed goodness-of-fit values differ significantly from those generated by comparing ordinations produced by the randomized permuted values. So, neither the goodness-of-fit statistics nor the p-values for either the Mantel Test and the PROTEST have a definitive value that can be considered a threshold between two ordinations that may be considered the same or different. Because the subsampled portion is derived from the complete dataset, there is almost by definition a degree of similarity between the two datasets, and so a high likelihood that they will produce low p-values.

Because there are no known thresholds between goodness-of-fit values (Mantel R-statistics and PROTEST $m^2$-values) that indicate a complete dataset result is the same as the subsampled dataset result, I conducted a visual comparison of 6,000 ordination pairs to estimate a threshold of goodness-of-fit values, i.e. at what goodness-of-fit value did I consistently interpret the two ordinations similarly? Only 6,000 were examined because it would be too time consuming to visually inspect all comparisons for all subsamples for all types of datasets (1,985,000 total). The datasets used for each of the 6,000 ordination comparisons were simulated in the same manner as the simulated-created datasets and thus constitute a representative sample. Three different methods of modifying that initial dataset were conducted.

The first modification, for the first 2,000 comparisons, was a 50% subsample of the initial dataset. The ordination of the initial dataset was compared with the ordination of the 50% subsample of that initial dataset. This was repeated 2,000 times, changing various parameters of

the simulated-created datasets. For the second modification method, I switched the position of one sample in the dataset. For example, sample 1 became sample 20, and all the remaining samples shifted position down by one (i.e., sample 2 became sample 1, and so on). If the samples were already arranged in discrete clusters, sample 1 would move to a different cluster for the Mantel Test and PROTEST analyses. For the third modification method, a "complementary environmental" dataset was created based on the same simulated-gradient as the primary dataset. This dataset modification represents a set of environmental or ecological variables to compare with the primary dataset. The two additional modifications were conducted to explore if different types of dataset comparison led to different thresholds.

For each of the 6,000 ordination comparisons, I gave a ranking of 1, 2, or 3 corresponding to the degree of similarity between ordinations, with rank 1 indicating that the two ordinations would be interpreted as having essentially the same result. Even though some researchers may disagree with whether or not two ordinations were the same (for those I ranked 1), any possible differences that may arise between our ranking and how other researchers may interpret the similarity of two ordination results would not likely change the relative patterns between the qualitative and quantitative comparison methods, especially averaged over 6,000 ordination comparisons.

A rank of 1 was given to two ordinations that were almost indistinguishable. Fewer than 50% of the sample points differ in position, and those sample points that do change position do not change the groupings or rank order of the samples. I assumed most researchers would agree that the two ordination results were the same and would lead to the same paleoecological interpretations Because I cannot be certain of how other researchers would assess and interpret ordination results, I was conservative in our ranking system. If there was doubt that researchers would interpret the two ordination results differently, I ranked the comparisons as a 2 or 3.

A rank of 2 was given if two or more samples differ in position radically through the ordination, if one cluster divided into multiple clusters, or if greater then 50% of sample points changed position. In the case of a rank of 2, the two ordinations may be considered different to some researchers.

A ranking of 3 was given if there was noticeable change in the position of many points. The points may move considerably along one or both ordination axes. In this case, the changes in the groupings and patterns of the points within the ordinations would clearly lead to different paleocommunity interpretations and conclusions.

2.6 Testing for differences between parameters

To examine if any variables had an effect on required sample sizes, I statistically compared (using either t-tests or ANOVAs) groups of goodness-of-fit statistics (Tables 6-3 and 6-4). All of the groups were gathered from the goodness-of-fit statistics that resulted from the comparison of datasets with a sample size of 50 to the complete sample size of 200. The

groupings of variables were based on richness (10, 20, and 50), evenness (low and high), and number of samples (5 and 10). By conducting these tests, I am able to determine if datasets with higher richness, evenness, or a greater number of samples may have influenced the goodness-of-fit between subsamples and complete datasets. I conducted these tests only at a sample size of 50 (and not 100, 50, 25 or 10) because this was the most similar value to the minimum required sample size of Forcino (2012) as well as most similar to the *post hoc* value determined herein (see below).

## 3. Results

### 3.1 Qualitative Assessment

The lowest values that indicated visual results that would be interpreted as the same (rank of 1) for the Mantel Test and PROTEST, were respectively $R = 0.86$ and $m^2 = 0.89$, at the 95% confidence interval (Figure 6-2). These values serve as estimates of the threshold between interpreting two datasets as different and the same. The 95% confidence interval of the rank of 2 would considerably lower the two threshold values, particularly because there was a much larger standard deviation. However, there was uncertainty if those comparisons given a rank of 2 would be interpreted as the same; using only a rank of 1 is more conservative.

### 3.2 Real datasets

With the exception of one dataset, the Mantel Test R-statistics were greater than the threshold of $R = 0.86$ for all sample sizes greater than 28 individuals (Figure 6-3a). When the median sample size is less than 28 individuals the R-statistics rapidly decrease. The one dataset that is below the threshold was Ryu et al. (2011), which contained primarily ostracods with an initial median sample size of 4939.

The Procrustean Randomization Test (PROTEST) $m^2$-values were not consistently above the threshold of $m^2 = 0.89$ (Figure 6-3b). The $m^2$-values were greater than 0.76 at median sample sizes greater than 58; the $m^2$-values then decrease rapidly.

### 3.3. Simulated-created datasets

The Mantel Test R-statistics are greater than the threshold of $R = 0.86$ for all sample sizes greater than 100 individuals (Figure 6-4a). The R-statistic values are consistently large as sample size decreases, until the median sample size is 54 individuals. When the median sample size is less than 54 individuals the R-statistics rapidly decrease. The PROTEST $m^2$-values are greater than the threshold of $R = 0.89$ for all sample sizes greater than 68 individuals (Figure 6-4b). The $m^2$-values above or slightly below the threshold value of 0.89 above a median sample size of 50 individuals. At a median sample size less than 50 individuals the $m^2$-values rapidly decrease. No pattern or separation in the goodness-of-fit statistics (both the Mantel Test R-statistics and the PROTEST $m^2$-values) was found among the variables: numbers of taxa, numbers of samples, or initial median sample size (Table 6-4).

### 3.4 Selected-created datasets

There was greater variation in the goodness-of-fit statistics among the selected-created dataset comparisons than the other two dataset types (Figure 6-5). There was no clear plateau or rapid decrease of goodness-of-fit statistics. Eighty-eight of the datasets were specifically constructed to have either low or high evenness. The 44 low evenness datasets had individual-sample evenness values ranging from 0.14 to 0.98 and a mean evenness of 0.58 (Pielou's J Evenness; Magurran 1998), and the 44 high evenness datasets with individual-sample evenness values from 0.18 to 1.0, with a mean of 0.79. Of these 88 datasets, the low evenness datasets consistently led to greater goodness-of-fit statistic values (Table 6-3). The mixed evenness datasets, those in which the dataset included some high-evenness samples and some low-evenness samples, produced the greatest goodness-of-fit statistics out of all selected-created datasets (Table 6-3; Figure 6-5). One difference between the PROTEST $m^2$-values and the Mantel Test R-statistics was that the mixed evenness datasets' goodness-of-fit values from the PROTEST were similar to those of the low evenness datasets (Figure 6-6b).

The ranges and significant differences among the different parameters (evenness and numbers and samples and taxa) vary depending on the parameters and between the Mantel Test and PROTEST (Table 6-3 and 6-4). Of the nine various dataset structures within each set of parameters, there is no consistent dataset structure that led to higher goodness-of-fit statistics.

## 4. Discussion

### 4.1 Sample size requirements

Both the 44 real datasets and the 220 simulated-created datasets demonstrate that smaller sample sizes produce the same multivariate, abundance-based community results as larger sample sizes, in the sense that the results are similar enough that they would be interpreted the same (Figures 6-3 and 6-4). Although there were some outliers below the respective estimated threshold values (the goodness-of-fit statistics separating results that would be interpreted as the same from those that would be interpreted as different), the vast majority of the subsample results were above the thresholds for sample sizes greater than 54 and 58 for the Mantel Test and PROTEST, respectively. Furthermore, the mean subsample goodness-of-fit statistics for all datasets are at or close to the threshold values ($R = 0.86$ and $m^2 = 0.89$) for sample sizes greater than 54 and 58, indicating that all the median sample sizes greater than these values produced the same results as larger sample sizes. The point at which the mean subsample goodness-of-fit statistics rapidly decrease is an additional metric with which to gauge when the subsamples do not produce the same results as larger sample sizes. As these results are based on median sample sizes within the dataset, a minimum sample-size of 58 individuals per sample would almost certainly be representative for use with these types of multivariate analyses, and as such, 58 individuals is a conservative recommendation for a minimum sample-size to be collected in the field. This sample size estimate is smaller than previous research that used different methods (i.e., probability estimates) to determine that 300 individuals per community

are required for ecological research (Phleger 1960, Chang 1967, Patterson and Fishbein 1989, Fatela and Taborda 2002, Choa et al. 2005, Schloss and Handelman 2007). However, the approach of comparing multivariate results is a more practical approach; most community research applies multivariate techniques. While a smaller sample size may not capture the exact diversity of a community, the smaller sample size would still maintain the general position and order of samples in ordination space as well as the identification of related groupings or gradients of communities.

Furthermore, a median sample size of 58 individuals is less than the median sample sizes of the 44 real datasets employed, which was 146 individuals with a range of 10 to 24,812. These 44 datasets are representative of community studies in the recent literature, and therefore, represent the range of typical median sample sizes collected by ecologists. Community studies can collect fewer individuals per sample and still obtain the same meaningful results.

I examined real datasets from a range of environments, geographic locations, and containing a range of taxonomic groups as well as both terrestrial and marine taxa (Figure 6-3; Table 6-1). In addition, the real and simulated-created datasets contain a range of numbers of samples, numbers of taxa, and evenness, resulting in datasets spanning an extremely broad range of possible communities. The results were consistent across this broad range of real or realistic communities. However, I did not examine methods of tallying taxa other than abundance counts and only multivariate analytical methods were used. So, if community researchers use methods other than abundance counts and multivariate statistics employed herein, the present study cannot provide insight into the sample size requirements.

Based on statistical significance, the majority of the comparisons using either the Mantel Test and PROTEST were significantly similar ($p < 0.01$). However, it is not clear whether a given Mantel Test or PROTEST goodness-of-fit value would actually indicate similarity of interpretation. The situation is somewhat analogous to a simple correlation where it is possible to have an extremely significant, but still weak correlation. There were numerous cases of $m^2$-values or R-statistics greater than 0.6 that were associated with p-values less than 0.01. This raises two important points: first, a visual assessment, and consequent identification of a threshold-value at which similarity is more certain, is necessary (Figure 6-2). In the majority of cases, I observed that pairs of ordination plots assigned a rank of 2, indicative of questionable similarity, would still have a p-value less than 0.01. The result may be significant but is not indicative of a strong similarity. Second, had I relied entirely on significance as a means of assessment, the general result would be that sample-sizes between ~ 20 and 30 individuals per sample would be deemed sufficient. Thus, the use of the visual assessment and derived $m^2$-values and R-statistics thresholds, and the resulting conclusion that the requisite minimum sample-size is approximately 58, is conservative.

4.2 When are larger sample sizes required?

The range of the goodness-of-fit statistics produced by the subsample comparisons of the 132 selected-created datasets was almost double the range of the goodness-of-fit statistics from the real and simulated-created datasets (Tables 6-3 and 6-4). Thus, there are circumstances where larger median sample sizes (> 58 individuals) are required for multivariate, abundance-based ecological research.

4.2.1 Evenness

Of the three parameters (number of samples, number of taxa, and evenness) that were systematically varied among the 132 created datasets, evenness had the greatest effect on whether or not the subsamples of a dataset produced the same multivariate result as the complete dataset. The low evenness datasets had consistently greater goodness-of-fit statistics than the high evenness datasets (Tables 6-3 and 6-4; Figure 6-5). There is a significant difference between the low and high evenness datasets for both the R-statistics and $m^2$-values (Table 6-4); datasets containing samples with consistently high evenness may require larger sample sizes.

However, the mixed evenness datasets have greater R-statistics than the low or high evenness datasets (Table 6-3). Although the low evenness datasets did have consistently greater R-statistics than the high evenness datasets, the range of low evenness R-statistics was still much greater than that of the mixed evenness datasets (Table 6-1). The PROTEST $m^2$-values for the mixed evenness datasets were not as consistently large as the Mantel R-statistics (Table 6-3; Figure 6-5b). Furthermore, the $m^2$-values for the mixed evenness datasets overlap with the low evenness dataset $m^2$-values.

A possible reason the mixed evenness datasets produced consistently greater R-statistics when compared with other dataset parameters (among the 132 selected-created datasets) is because of the two different groups of samples within each dataset—one group with higher evenness and one group with lower evenness. This produces a strong gradient or two clusters of communities, based on dissimilarity measures that are strongly different from one another. Such strong within-dataset differences manifest even with smaller sample sizes; the strong difference in evenness between groups of samples was clear even when fewer individuals per sample are collected. It should be noted that the mixed evenness datasets are probably much more likely to be found in nature than the extreme cases of uniformly high- or low-evenness across communities. While consistently high evenness communities may require larger sample sizes, such systems would appear to be rare in nature; none of the real datasets used in this study had mean evenness values as great as those of the selected-created high evenness datasets.

4.2.2 Number of taxa

There were differences found between the different numbers of taxa, specifically datasets with more taxa leading to lower goodness-of-fit statistics (Tables 6-3 and 6-4). However, these differences were not consistent throughout all numbers of taxa and the two comparison methods (Tables 6-3 and 6-4). Overall, this effect of the number of taxa on the

required sample size is minor relative to the complete analysis of all 396 datasets. In addition to the effect of the number of taxa on selected-created dataset, eleven of the real datasets and 50 of the simulated-created datasets had more than 50 taxa, and all of these datasets produced consistently high goodness-of-fit statistics between subsampled and corresponding complete datasets (Tables 6-3 and 6-4). So, the majority of datasets, even those with a larger number of taxa, still demonstrate that smaller samples sizes are sufficient for multivariate community research.

4.2.3 Number of samples

There was a significant difference found between selected-created datasets with 5 and 10 samples (Tables 6-3 and 6-4; Figure 6-5). The datasets were constructed so that each sample within the selected-created datasets represents a different community. With a greater number of communities (10 versus 5), the multivariate analysis is more likely to distinguish between most of the communities even with fewer individuals per sample. This is similar to the mixed evenness datasets; the multivariate analyses are better able to distinguish between the two dichotomous groups of samples even at smaller sample sizes, which was likely the cause for the high goodness-of-fit statistics. When the datasets are limited to 5 communities, there is less of a chance that the community gradient will still be apparent when sample sizes decrease. When there are 10 communities, there is a higher probability that one or two communities will remain intact even at smaller sample sizes, producing the community gradient in ordination space. This is additional evidence that homogeneity of communities within a dataset may require larger sample sizes. However, it should be noted that community researchers are often attempting to find out what is driving community structure or community change, and many, if not most, studies deliberately sample along suspected gradients or between environmental conditions known to be different. Environmental homogeneity among sampled communities is not a common goal. So, this issue of larger sample size requirements among homogeneous communities should not have a grave impact on community research.

5. Conclusion

The primary goal of this study was to determine if smaller sample sizes produce the same results as larger, more typically collected sample sizes. Examining 44 real datasets and 220 simulated-created datasets, I found evidence that smaller sample sizes (i.e., 58 individuals) produce the same community results as larger sample sizes. To detect possible dataset parameters that require larger sample sizes, I subsampled 132 selected-created datasets in which the number of samples, number of taxa, and evenness were systematically altered to test for an effect on the required sample size. I found evenness strongly influenced the goodness-of-fit between subsamples and corresponding complete datasets; high evenness datasets produced lower goodness-of-fit statistics than low evenness and mixed evenness datasets. Although high evenness datasets may have led to lower a goodness-of-fit, few community studies would consist

entirely of uniformly high-evenness communities. In addition, the number of taxa had a minor influence on the goodness-of-fit between subsamples and corresponding complete datasets; datasets with more taxa produced lower goodness-of-fit statistics than datasets with fewer taxa. However, this effect caused by the number of taxa was not consistent among all 396 datasets, and indeed many of the taxonomically richest datasets had very large goodness-of-fit values. Also, homogeneity among communities within a dataset produced lower goodness-of-fit statistics. This finding that smaller sample sizes are sufficient for multivariate community research could fundamentally change sampling protocols for community studies; smaller sample-sizes would save resources and enable researchers to collect more samples instead of fewer, larger samples. In addition, sample size-limited communities, which might previously have been ignored, can be used with statistical confidence that the research is using a representative sample.

## 6. Acknowledgements

This research would not have been possible without the NSF- EAR-0746072 grant to Leighton. Thank you to JC Cahill, Justine Karst, Tan Bao, Jon Bennett, and the rest of the University of Alberta Plant Ecology Group, without whom this research would lack the proper ecological edge it required. I would like to thank Chris Schneider, Emily Stafford, and Brian Chatterton for their feedback on the manuscript and the project overall.

## 7. Bibliography

Anderson, J., L. Vermeire, and P. B. Adler. 2011. Fourteen years of mapped, permanent quadrats in a northern mixed prairie, USA. Ecology 92:1703.

Arai, H. P., and D. R. Mudry. 1983. Protozoan and metazoan parasites of fishes from the headwaters of the Parsnip and McGregor Rivers, British Columbia: a study of possible parasite transfaunations. Canadian Journal of Fisheries and Aquatic Sciences 40:1676–1684.

Arthur, J. R., L. Margolis, and H. P. Arai. 1976. Parasites of fishes of Aishihik and Stevens Lakes, Yukon Territory, and potential consequences of their interlake transfer through a proposed water diversion for hydroelectrical purposes. Journal of the Fisheries Research Board of Canada 33:2489–2499.

Bangham, R. V. 1955. Studies on fish parasites of Lake Huron and Manitoulin Island. American Midland Naturalist 53:184–194.

Beehler, B. 1983. Frugivory and polygamy in birds of paradise. The Auk 100:1–12

Bennington, J. B., and S. D. Rutherford. 1999. Precision and reliability in paleocommunity comparisons based on cluster-confidence intervals: how to get more statistical bang for your statistical buck. PALAIOS 14:506–515.

Bush, A. M., and R. I. Brame. 2010. Multiple paleoecological controls on the composition of marine fossil assemblages from the Frasnian (Late Devonian) of Virginia, with a comparison of ordination methods. Paleobiology 36:573–591.

Cao Y., D. D. Williams, and D. P. Larsen. 2002. Comparison of ecological communities: the problem of sample representativeness. Ecological Monographs 72:41–56.

Causse, R., C. Ozouf-Costaz, P. Koubbi, D. Lamy, M. Eleaume, A. Dettaı, G. Duhamel, F. Busson, P. Pruvost, A. Post, R. J. Beaman, and M. J. Riddle. 2011. Demersal ichthyofaunal shelf communities from the Dumont d'Urville Sea (East Antarctica). Polar Science 5:272–285.

Chang, Y. M. 1967. Accuracy of fossil percentage estimation. Journal of Paleontology 41:500–502.

Chao, A., R. L. Chazdon, R. J. Colwell, and T. J. Shen. 2005. A new statistical approach for assessing similarity of species composition with incidence and abundance data. Ecology Letters 8:145–159.

Chinniah, V. C., and W. Threlfall. 1978. Metazoan parasites of fish from the Smallwood Reservoir, Labrador, Canada. Journal of Fish Biology 13:203–213.

Clarke, K. R., and M. Ainsworth. 1993. A method and linking multivariate community structure to environmental variables. Marine Ecology 92:205–219.

Dahle, S., S. G. Denisenko, N. V. Denisenko and S. J. Cochrane. 1998. Benthic fauna in the Pechora Sea. Sarsia 83:183–210.

Dechtiar, A. O. 1972. Parasites of fish from Lake of the Woods, Ontario. Journal of Fisheries Research Board of Canada 29:275–283.

Fall, L. M., and T. D. Olszewski. 2010. Environmental disruptions influence taxonomic composition of brachiopod paleocommunities in the Middle Permian Bell Canyon Formation (Delaware Basin, West Texas). PALAIOS 25:247–259.

Fatela, F., and R. Taborda. 2002. Confidence limits of species proportions in microfossil assemblages. Marine Micropaleontology 45:169–174.

Forcino, F. L. 2012. Multivariate assessment of the required sample size for paleocommunity research. Palaeogeography, Palaeoclimatology, Palaeoecology 315:134–141.

Forcino, F. L., E. S. Stafford, J. J. Warner, A. E. Webb, L. R. Leighton, C. L. Schneider, T. S. Michlin, L. M. Palazzolo, J. R. Morrow, and S. A. Schellenberg. 2010. Effects of data categorizations on paleocommunity analysis: a case study from the Pennsylvanian Finis Shale of Texas. PALAIOS 25:144–157.

Goff, F. G., and R. Mitchell. 1975. A Comparison of species ordination results from plot and stand data. Vegetatio 31:15–22.

Hayek, L. C., and M. A. Buzas. 1997. Surveying Natural Populations. Columbia University Press, New York.

Hughes, R. N., and M. I. H. Thomas. 1971. The classification and ordination of shallow-water benthic samples from Prince Edward Island, Canada. Journal of Experimental Marine Biology and Ecology 7:l–39.

Ieno, E. N., and R. O. Bastida. 1998. Spatial and Temporal Patterns in Coastal Macrobenthos of Samborombon Bay, Argentina: A Case Study of Very Low Diversity. Estuaries 41:690–699.

Jackson, D. A. 1995. PROTEST: A Procrustean randomization test of community environment concordance. Écoscience 2:297–303.

Jalonen, J., I. Vanha-Majamaa, and T. Tonteri. 1998. Optimal sampling and plot size for inventory of field and ground layer vegetation in mature *Myrtillus*-type boreal spruce forest. Annales Botinici Fennici 35:191–196.

Kang, S., and A. L. Mills. 2006. The effect of sample size in studies of soil microbial community structure. Journal of Microbiological Methods 66:242–250.

Kinnunen, H., and J. Tiainen. 1999. Carabid distribution in a farmland mosaic: the effect of patch type and location. Ann. Zool. Fennici 36: 149–158.

Kruskal, J. B. 1964. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychomerika 29:1–27.

Legendre P., and L. Legendre. 1998. Numerical Ecology. Elsevier, Amsterdam, The Netherlands.

Leong, T. S., and J. C. Holmes. 1981. Communities of metazoan parasites in open water fishes of Cold Lake, Alberta. Journal of Fish Biology 18:693–713.

Mantel, N. 1967. The detection of disease clustering and a generalized regression approach. Cancer Research 27:209–220.

McCune, B., and J. B. Grace. 2002. Analysis of Ecological Communities. MjM Software Design, Gleneden Beach, Oregon.

Millar, R.B., M. J. Anderson, and N. Tolimieri. 2011. Much ado about nothings: using zero similarity points in distance-decay curves. Ecology 92:1717–1722.

Nicolaidou, A., K. Petrou, K. A. Kormas and S. Reizopouloz. 2006. Inter-annual variability of soft bottom macrofaunal communities in two Ionian Sea lagoons. Hydrobiologia 555:89–98.

Oksanen, J., F. G. Blanchet, R. Kindt, P. Legendre, R. B. O'Hara, G. L. Simpson, P. Solymos, M. H. H. Stevens, and H. Wagner. 2012. vegan: Community Ecology Package. R package version 2.14, http://CRAN.R-project.org/package=vegan.

Otypkova, Z., and M. Chytry. 2006. Effects of plot size on the ordination of vegetation samples. Journal of Vegetation Science 17:465–472.

Patterson, R. T., T. and Fishbein. 1989. Re-examination of the statistical methods used to determine the number of point counts needed for micropaleontological quantitative research. Journal of Paleontology 63:245–248.

Peres, C. A. 1997. Primate community structure at twenty western Amazonian flooded and unflooded forest .Journal of Tropical Ecology. 13: 381–405.

Peres-Neto, P. R., and D. A. Jackson. 2001. How well do multivariate datasets match? The advantages of a Procrustean superimposition approach over the Mantel Test. Oecologia

129:169–178.

Petraitis, P.S., H. Liu, and E. C. Rhile. 2009. Barnacle, fucoid, and mussel recruitment in the Gulf of Maine, USA, from 1997 to 2007. Ecology 90:571.

Phleger, F. B. 1960. Ecology and Distribution of Recent Forminifera. Johns Hopkins Press, Baltimore, MD.

R Development Core Team. 2012. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Ramesh, B. R., M. H. Swaminath, S. V. Patil, Dasappa, R. Pélissier, P. D. Venugopal, S. Aravajy, C. Elouard, and S. Ramalingam. 2010. Forest stand structure and composition in 96 sites along environmental gradients in the central Western Ghats of India. Ecology 91:3118–3118.

Ranjard, L., D. P. H. Lejon, C. Mougel, L. Schehrer, D. Merdinoglu, and R. Chaussod. 2003. Sampling strategy in molecular microbial ecology: influence of soil sample size on DNA finger printing analysis of fungal and bacterial communities. Environmental Microbiology 5:1111–1120.

Repecka, R., and E. Mileriene. 1991. Species composition and abundance in shore-zone fish communities of the Kursiu bay during 1985-1989. Elologija 4:68–80.

Ryu, J., Khim, J.S., Kang, S.G., Kang, D., Lee, C., and Koh, C. 2011. The impact of heavy metal pollution gradients in sediment on benthic macrofauna at population and community levels. Environmental Pollution 159:2622-2629.

Schloss, P. D. 2008. Evaluating different approaches that test whether microbial communities have the same structure. The ISME Journal 2:265–275.

Schloss, P. D., and J. Handelman. 2007. The last word: books as a statistical metaphor for microbial communities. Annual Review of Microbiology 61:23–34.

Skłodowski, J. J. W., and J. Porowski. 2000. Composition and structure of carabid communities (Col., Carabidae) living in small mid-field pine stands. Sylwan Vol. 144:45–64.

Snow, B.K., and D.W. Snow. 1971. The feeding ecology of Tanagers and honeycreeperes in Trinidad. Auk 88: 291–322.

Snow, B.K., and D.W. Snow. 1988. Birds and Berries. Calton, England.

Stevens, C. J., C. Dupré, E. Dorland, C. Gaudnik, D. J. G. Gowing, M. Diekmann, D. Alard, R. Bobbink, E. Corcket, J. O. Mountford, V. Vandvik, P. A. Aarrestad, S. Muller, and N. B. Dise. 2011. Grassland species composition and biogeochemistry in 153 sites along environmental gradients in Europe. Ecology 92:1544–1544.

Ulrich, W., and Gotelli, N.J. 2010. Null model analysis of species associations using abundance data. Ecology 91:3384-3397.

Ulrich, W., and M. Zalewski. 2006. Abundance and co-occurrence patterns of core and satellite species of ground beetles on small lake islands. Oikos 114:338–348.

VanNimwegen, R. E., J. Kretzer, and J. F. Cully. 2008. Ecosystem engineering by a colonial mammal: how prairie dogs structure rodent communities. Ecology 89:3298–3305.

Warwick, R. M., and K. R. Clarke. 1993. Comparing the severity of disturbance: a meta-analysis of marine macrobenthic community data. Marine Ecology Progress Series 92:221–231.

Wong, D. C. L., L. Maltby, D. Whittle, P. Warren, and P. B. Dorna. 2004. Spatial and temporal variability in the structure of invertebrate assemblages in control stream mesocosms. Water Research 38:128–138.

**Table 6-1**. Characteristic of the 44 real datasets (some of the citations contain multiple datasets).

| Citation | Median Sample Size | Number of Samples | Number of Taxa | Mean Evenness | Environment | Primary Taxonomic Group | Geographic Location |
|---|---|---|---|---|---|---|---|
| Beehler 1983 | 97 | 8 | 31 | 0.72 | Forest | Birds and plants | Papua New Guinea |
| Arthur et al. 1976 | 85 | 38 | 17 | 0.74 | Lake | Parasites | Yukon, Canada |
| Cause et al. 2011 | 20 | 43 | 53 | 0.74 | Subtidal marine | Parasites | Dumont d'Urville Sea (East Antarctica) |
| Wong et al. 2004 | 24812 | 12 | 13 | 0.46 | Fresh water streams | Invertebrates | Kent, Uk, and Mississippi, USA |
| VanNimwegen et al. 2008 | 75 | 4 | 7 | 0.69 | Grasslands | Prairie dogs | Kansas, USA |
| Ieno and Bastido 1998 | 853 | 7 | 13 | 0.75 | Benthic marine | Bivalves and ploychaetes | Samborombon Bay, Argentina |
| Kinnunen and Tiainen 1999 | 147 | 40 | 7 | 0.59 | Farmland | Beetles | Finland |
| Nicolaidou et al. 2006 | 890 | 18 | 48 | 0.64 | Benthic lagoon | Bivalves | Ionian Sea, Greece |
| Arai and Budry 1983 | 114 | 17 | 53 | 0.83 | River | Fish and parasites | British Columbia, Canada |
| Peres 1997 | 110 | 12 | 12 | 0.94 | Forest | Primates | Brazil |
| Dahle et al. 1998 | 944 | 15 | 421 | 0.70 | Benthic brackish | Marine invertebrates | Pechora Sea, Russia |
| Repecka and Mileriene 1991 | 511 | 19 | 20 | 0.95 | Marine | Fish | Kursia Bay, Lithuania |
| Hughes and Thomas 1971 | 94 | 16 | 16 | 0.69 | Benthic Estuary | Bivalves | Prince Edward Island, Canada |
| Hughes and Thomas 1971 | 76 | 21 | 18 | 0.67 | Benthic Estuary | Bivalves | Prince Edward Island, Canada |
| Hughes and Thomas 1971 | 235.5 | 14 | 14 | 0.51 | Benthic Estuary | Bivalves | Prince Edward Island, Canada |
| Hughes and Thomas 1971 | 648 | 51 | 51 | 0.72 | Benthic Estuary | Bivalves | Prince Edward Island, Canada |
| Ryu et al. 2011 | 4939 | 7 | 36 | 0.53 | Benthic marine to | Benthic animals | Incheon North Harbor, |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | | | brackish | | Korea |
| Skrodowski and Porowski 2000 | 210 | 25 | 22 | 0.73 | Pine forest | Beetles | Poland |
| Snow and Snow 1971 | 146 | 13 | 65 | 0.76 | Neotropical forest | Birds | Trinadad |
| Snow and Snow 1988 | 234 | 7 | 12 | 0.50 | Mixed terrestrial | Birds and plants | England |
| Snow and Snow 1971 | 1674 | 9 | 35 | 0.70 | Neotropical forest | Birds | Trinadad |
| Ulrich and Zalewski 2006 | 145 | 11 | 17 | 0.76 | Lake Islands | Beetles | Multiple |
| Dechitar 1972 | 338 | 31 | 144 | 0.93 | Lake | Parasites | Ontario, Canada |
| Anderson et al. 2011 | 850.5 | 42 | 39 | 0.52 | Northern mixed prairie | Grassland plants | Montana, USA |
| Anderson et al. 2011 | 9.5 | 10 | 6 | 0.80 | Northern mixed prairie | Grassland plants | Montana, USA |
| Anderson et al. 2011 | 261 | 25 | 15 | 0.50 | Northern mixed prairie | Grassland plants | Montana, USA |
| Anderson et al. 2011 | 52 | 29 | 14 | 0.50 | Northern mixed prairie | Grassland plants | Montana, USA |
| Anderson et al. 2011 | 29 | 42 | 31 | 0.55 | Northern mixed prairie | Grassland plants | Montana, USA |
| Anderson et al. 2011 | 35 | 39 | 17 | 0.44 | Northern mixed prairie | Grassland plants | Montana, USA |
| Anderson et al. 2011 | 29 | 42 | 41 | 0.63 | Northern mixed prairie | Grassland plants | Montana, USA |
| Anderson et al. 2011 | 118 | 37 | 46 | 0.50 | Northern mixed prairie | Grassland plants | Montana, USA |
| Anderson et al. 2011 | 30 | 37 | 43 | 0.67 | Northern mixed prairie | Grassland plants | Montana, USA |
| Anderson et al. 2011 | 248 | 41 | 46 | 0.60 | Northern mixed prairie | Grassland plants | Montana, USA |
| Anderson et al. 2011 | 573 | 42 | 37 | 0.47 | Northern mixed prairie | Grassland plants | Montana, USA |
| Anderson et al. 2011 | 53 | 41 | 27 | 0.71 | Northern mixed prairie | Grassland plants | Montana, USA |
| Anderson et al. 2011 | 20 | 41 | 30 | 0.72 | Northern mixed prairie | Grassland plants | Montana, USA |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Miller et al. 2011 | 6431 | 68 | 117 | 0.66 | Marine | Fish | Pacific coast, USA |
| Petraitis et al. 2009 | 301 | 60 | 3 | 0.67 | Intertidal | Bivalves and algae | Maine, USA |
| Ramesh et al. 2010 | 132 | 95 | 334 | 0.77 | Tropical terrestrial | Plants | Karnataka, India |
| Stevens et al. 2011 | 33 | 280 | 155 | 0.90 | Grasslands | Plants and bryophytes | Atlantic coast, Europe |
| Stevens et al. 2011 | 51 | 40 | 100 | 0.93 | Grasslands | Plants and bryophytes | Atlantic coast, Europe |
| Stevens et al. 2011 | 52 | 445 | 355 | 0.95 | Grasslands | Plants and bryophytes | Atlantic coast, Europe |
| Ulrich and Gotelli 2010 | 248 | 6 | 25 | 0.77 | River | Fish | British Columbia, Canada |
| Ulrich and Gotelli 2010 | 495 | 8 | 99 | 0.88 | Lake Islands | Beetles | Multiple |

**Table 6-2**. Characteristics of the 220 simulated created dataset

| Datasets | Number of samples | Number of taxa | Gradient Size | Median sample size |
|---|---|---|---|---|
| 1-10 | 15 | 15 | 1000 | 8418 |
| 11-20 | 15 | 15 | 5000 | 41528 |
| 21-30 | 20 | 20 | 100 | 178 |
| 31-40 | 20 | 20 | 5000 | 58014 |
| 41-50 | 20 | 30 | 100 | 278 |
| 51-60 | 20 | 40 | 100 | 256 |
| 61-70 | 20 | 40 | 100 | 356 |
| 71-80 | 20 | 50 | 100 | 255 |
| 81-90 | 20 | 50 | 100 | 493 |
| 91-100 | 25 | 100 | 100 | 936 |
| 101-110 | 30 | 20 | 100 | 194 |
| 111-120 | 30 | 60 | 100 | 376 |
| 121-130 | 40 | 20 | 100 | 169 |
| 131-140 | 50 | 20 | 100 | 181 |
| 141-150 | 50 | 50 | 100 | 474 |
| 151-160 | 50 | 50 | 5000 | 151434 |
| 161-170 | 50 | 75 | 100 | 761 |
| 171-180 | 50 | 100 | 100 | 982 |
| 181-190 | 50 | 200 | 100 | 2037 |
| 191-200 | 75 | 50 | 100 | 484 |
| 201-210 | 100 | 50 | 100 | 468 |
| 211-220 | 200 | 50 | 100 | 462 |

**Table 6-3**. Minimum and maximum goodness-of-fit statistics for the different parameters at a sample size of 50 for the selected-created datasets

| | Mantel Test | | PROTEST | |
|---|---|---|---|---|
| | Min | Max | Min | Max |
| All 132 | 0.53 | 0.98 | 0.56 | 0.98 |
| High Evenness | 0.53 | 0.92 | 0.65 | 0.98 |
| Low Evenness | 0.74 | 0.97 | 0.56 | 0.91 |
| Mixed Evenness | 0.92 | 0.98 | 0.62 | 0.98 |
| 5 Samples | 0.53 | 0.98 | 0.70 | 0.98 |
| 10 Samples | 0.64 | 0.97 | 0.56 | 0.98 |
| Richness = 10 | 0.58 | 0.98 | 0.65 | 0.98 |
| Richness = 20 | 0.53 | 0.97 | 0.56 | 0.97 |
| Richness = 50 | 0.59 | 0.97 | 0.75 | 0.97 |
| Richness = 10 (mixed evenness) | 0.92 | 0.98 | 0.65 | 0.98 |
| Richness = 50 (mixed evenness) | 0.92 | 0.97 | 0.89 | 0.98 |

**Table 6-4**. Tests for significant differences in groupings of goodness-of-fit statistics at a sample size 50 for the selected-created datasets

| Type of test | Groups being tested | p-value |
|---|---|---|
| T-test | High and low evenness dataset R-statistics | $p < 0.001$ |
| T-test | High and low evenness datasets $m^2$-values | $p = 0.003$ |
| T-test | R-statistics for the datasets with 5 samples and 10 sample | $p = 0.03$ |
| T-test | $m^2$-values for the datasets with 5 samples and 10 sample | $p < 0.001$ |
| ANOVA | R-statistics for the datasets with a richness of 10, 20, and 50 | $p = 0.006$ |
| Bonferroni corrected T-test | R-statistics for the datasets with a richness of 20 and 50 | $p = 0.009$ |
| Bonferroni corrected T-test | R-statistics for the datasets with a richness of 10 and 20 | $p = 0.053$ |
| Bonferroni corrected T-test | R-statistics for the datasets with a richness of 10 and 50 | $p = 0.94$ |
| ANOVA | $m^2$-values for the datasets with a richness of 10, 20, and 50 | $p = 0.53$ |
| T-test | R-statistics the datasets with a richness of 10 and 50 (mixed evenness datasets) | $p = 0.09$ |
| T-test | $m^2$-values the datasets with a richness of 10 and 50 (mixed evenness datasets) | $p = 0.04$ |

**Figure 6-1**. Simplified visual representation of the dataset simulation process using 5 taxa and 5 samples.

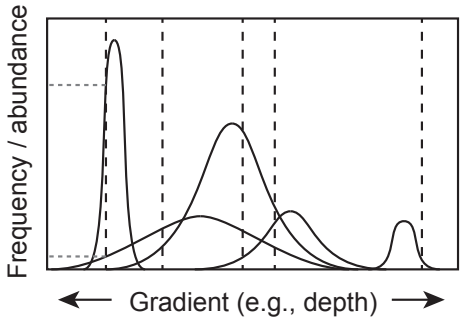


**Figure 6-2**. Box plots of the (a) Mantel Test R-statistics and (b) PROTEST $m^2$-values for the three qualitative visual rankings of the 6,000 ordination comparisons.

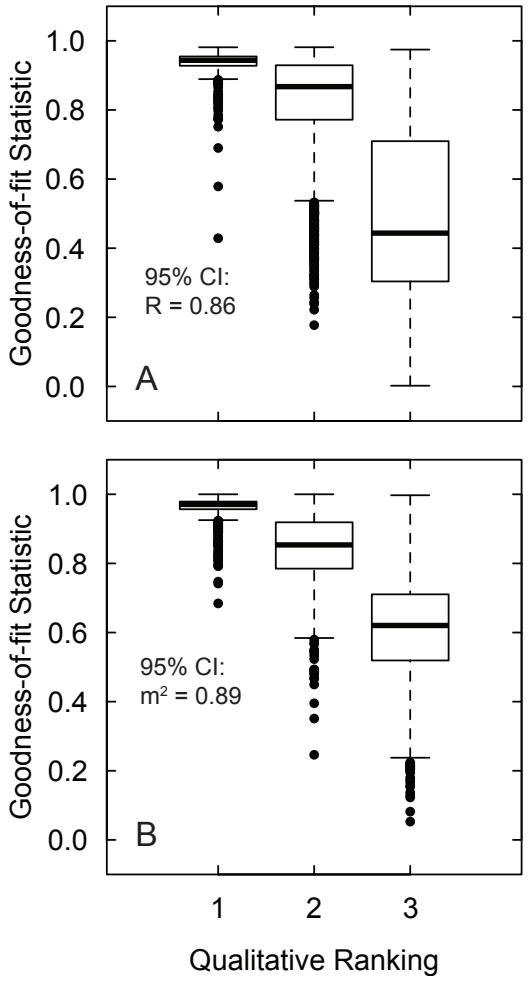


**Figure 6-3**. Box plots of the (a) Mantel Test R-statistics and (b) PROTEST $m^2$-values for the three qualitative visual rankings of the 6,000 ordination comparisons.

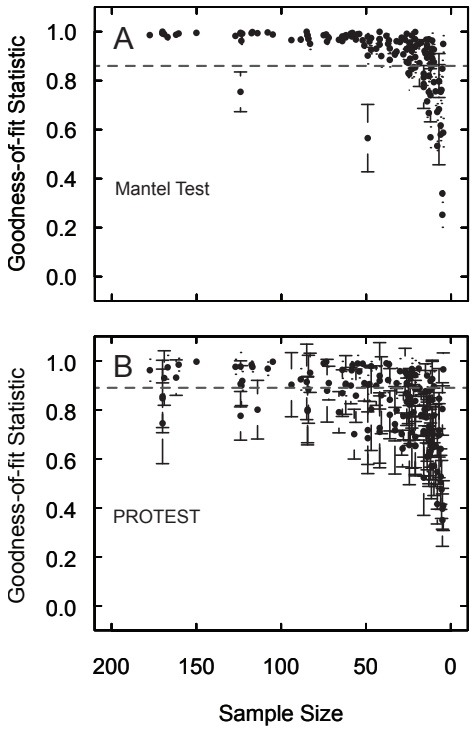


**Figure 6-4**. Mean (black circles) and standard deviations (black confidence lines) of the (a) Mantel Test R-statistics and (b) PROTEST $m^2$-values for the 1000 subsamples for each subsample size for the 220 simulated-created datasets plotted versus median sample size. The horizontal dashed grey line is the goodness-of-fit threshold value determined from the qualitative visual assessment.

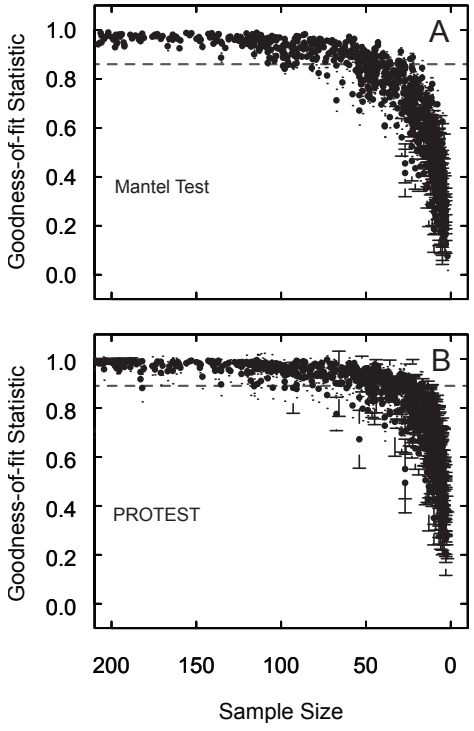
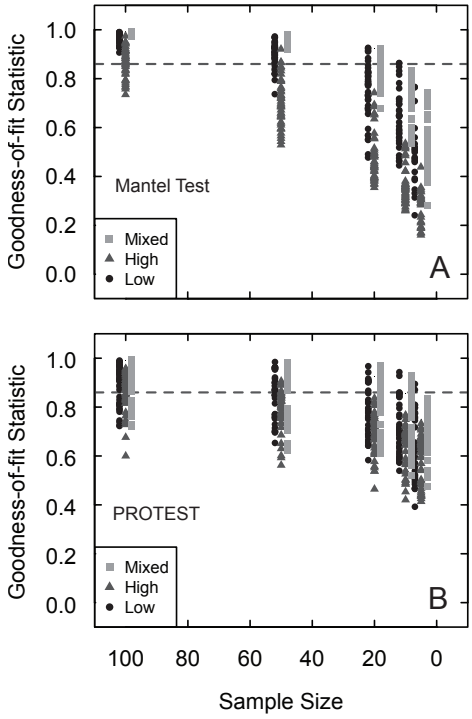


**Figure 6-5**. Mean of the (a) Mantel Test R-statistics and (b) PROTEST $m^2$-values for the 1000 subsamples for each subsample size for the 132 selected-created datasets plotted versus median sample size. The horizontal dashed grey line is the goodness-of-fit threshold value determined from the qualitative visual assessment.



## Chapter 7. Conclusions

The overall goal of the research of this thesis was to determine a standardized protocol for paleocommunity research. I attempted this by comparing the results and interpretations of different methods using a series of real and simulated community datasets. Although this approach is the most practical, it still can not be all encompassing to include all types of data and types of questions that may be of interest to a researcher. In other words, there may always be exceptions to a "best" paleocommunity set of methods. With that in mind, it is still important to determine if a certain set of methods produces a more accurate picture of past ecosystems or if there are certain techniques that skew results in one direction or another.

Of the five different methodological comparisons examined (the spatial and temporal resolution of sample collection, counting methods (abundance or biomass), the groups of organisms that should be examined, the taxonomic level of identification, and sample size), varying levels of certainty can be ascribed to the accuracy of the results produced. Below I discuss a few concluding remarks about each of these five different comparisons ordered from strongest evidence to weakest.

1. Sample size

Chapters 5 and 6 entail subsampling of 426 community datasets (real and simulated, modern and fossil) with the goal of examining whether smaller sample sizes produce the same multivariate statistical results as larger sample sizes. Although there are a few exceptions, sample sizes of 50 to 58 individuals (for fossil and modern datasets, respectively) generally produce the same multivariate community results as all larger sample sizes. This sample size of 50 to 58 is lower than previous research that determined 300 or more individuals per sample are required. However, these previous studies used completely different methods. Here, a more practical approach is taken by comparing real statistical results of real communities.

This finding that smaller sample sizes are sufficient for multivariate community research could fundamentally change sampling protocols for community studies; smaller sample sizes would save resources and enable researchers to collect more samples instead of fewer, larger samples. In addition, sample size-limited communities, which might previously have been ignored, can be used with statistical confidence that the research is using a representative sample.

2. Taxonomic level

Chapter 3 contained a meta-analysis of 28 datasets in which the multivariate statistical results of each dataset tallied at various taxonomic levels were compared. Family-level identification of specimens can lead to the same paleocommunity conclusions as genus-level identification, but inconsistencies generate enough uncertainty that paleocommunity research would benefit from genus-level identification of specimens. Due to the moderate-to-low goodness-of-fit statistics between genus-order and genus-class comparisons of paleocommunities, as well as the clear differences found in the qualitative-visual comparisons,

order and class did not reliably reproduce genus-level results. So, with moderate to high certainty, paleocommunity research should be conducted at the genus or species level.

3. Counting Methods

The two means of tallying fossil material examined here, abundance and point counts, produced different multivariate statistical results. However, previous research has demonstrated that abundance and biomass proxies can produce the same results. Although both methods may produce accurate representations of paleocommunities when using multivariate techniques, there may be some cases where one method or the other is more appropriate. So, whenever possible, I recommend examining both abundance and point counts (or some other biomass proxy) when conducting research on paleocommunities.

Additional data will shed more definitive light on which method may be most accurate, and which method may be better suited for a particular study. Hopefully, if this recommendation of using multiple counting methods is applied to future paleocommunity research efforts, additional case studies can be compiled. With an ample number of case studies, a more definitive conclusion about when to use abundance and when to use point counts can be provided.

4. Including all taxa versus brachiopod-only

Single taxonomic groups are often used for examining ecological and environmental variation through time. Here, I found different multivariate statistical results between the use of all taxa and brachiopod taxa only. Although more research is warranted prior to any definitive conclusions being stated, the evidence here demonstrates that taxa other than brachiopods can influence the community structure of a dataset. So, I recommend examining the greatest possible range of organisms that can be collected when conducting paleocommunity research.

5. Spatial versus temporal sampling resolution

I found that a minor variation in temporal scale led to a major change in the primary factor driving multivariate statistical results of paleocommunities. This suggests that patterns revealed at different scales are extremely sensitive to the temporal extent of sampling. This provides evidence that increasing the number of lateral samples per stratigraphic unit does not increase the accuracy of results when larger temporal scales are studied. Thus, when examining coarser-scale community variation, sampling effort is better-spent collecting samples from a greater number of stratigraphic units rather than replicating samples laterally.

Although further work and additional case studies on this hypothesis are definitely in order, this serves as evidence that there is some temporal scale at which the temporal community signal will overwhelm that of the spatial scale. Thus, since the question of interest dictates the temporal and spatial scale of the study, if the question requires a broader temporal scale, the problem of lateral variation is probably not as serious as has been thought; community paleoecology examinations at larger scales require fewer samples per stratigraphic horizon, bed, or unit. If further evidence supports this conclusion, this could save a great deal of resources and

time for paleontology researchers.