

REDUNDANCY IN DATA FROM STREAM SURVEYS

ROGER L. KAESLER*

Department of Geology, University of Kansas, Lawrence, Kansas 66045, U.S.A.

JOHN CAIRNS, Jr. and JOHN S. CROSSMAN†

Department of Biology and Center for Environmental Studies,
Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061, U.S.A.

(Received 12 December 1973)

Abstract—Data on the distribution of the biota of streams may be unnecessarily expensive and time-consuming to collect because of redundancy of information provided by various groups of organisms. Using correlation coefficients computed between corresponding elements of Q -mode similarity matrices, it was found that in the upper Potomac River aquatic insects gave the best agreement with the total biota of any subset tested ($r=0.603$). In the Clinch River, similarities among stations based on occurrences of Ephemeroptera and Coleoptera agreed most closely with similarities based on the total insect fauna, but the two orders were not closely similar to each other. This result suggests that study of both groups may reveal most of the information provided by the total insect fauna. Associations among stations based on occurrences of Gastropoda were closely similar to those based on the total noninsect macroinvertebrate fauna.

INTRODUCTION

Use of the biota to assess the health of aquatic environments of streams and the extent of their recovery from pollutional stress is a time-honored practice. Its success follows from the fact that the organisms inhabiting a stream form an intracting community with multiple cause-and-effect pathways. The biota provides information about the state of the environment that cannot be obtained from study of physical and chemical parameters alone, making the two types of information complementary rather than mutually exclusive. Most importantly, the biota responds through time to the collective impact of the various environmental parameters and thus acts as an information integrator. However, the significance of lists of formal taxonomic names is often difficult for non-biologists to appreciate and in large doses is undigestible even for professional biologists.

Several means of summarizing data have been proposed so that statements about the health of the aquatic environment need not be based on interpretation of faunal lists alone. Among these are (1) the target species approach (Mount, 1969; see also Cairns, *et al.*, 1972); (2) the saprobian system (Kolkwitz and Marsson, 1908; Gaufin and Tarzwell, 1952 and 1956; Beck, 1954 and 1955; Hynes, 1962; and Fjordingstad, 1965); (3) species

diversity (Margalef, 1958; Wilhm and Dorris, 1968; Buzas, 1972); (4) Patrick's method (Patrick, 1949); (5) the sequential comparison index (Cairns, *et al.*, 1968 and 1970; Cairns and Dickson, 1971; see also Simmons 1972); and (6) cluster analysis (Kaesler, 1966; Brown, 1969; Williams, 1971; Kaesler and Cairns, 1972; Wiebe, 1972; Crossman, Cairns, and Kaesler, in press; Kaesler, Crossman, *et al.*, in press).

The suggestion has been made elsewhere that many of the data collected during surveys of streams are redundant, so that nearly as much information could be obtained by studying fewer groups of organisms than are now studied (Kaesler and Cairns, 1972; Crossman, *et al.*, in press). If so, the cost of such surveys could be greatly reduced, and the rate of return of information could be increased. Moreover, if only a small subset of the aquatic biota or of the benthic macroinvertebrate community were to be chosen for study during such surveys, the degree of accuracy of identification and discrimination of organisms would probably increase because the ecologists making the identifications could become more familiar with the taxonomy of the group of organisms selected for study.

The purpose of this research is to examine the amount of redundancy that resides in data from several surveys of streams and to suggest means of eliminating the redundancy.

METHODS

The methods of cluster analysis we have used previously (Kaesler and Cairns, 1972 and other papers cited

* Done while visiting Professor with Department of Biology, Virginia Polytechnic Institute and State University.

† Presently employed by Teledyne-Brown Engineering, Research Park, Mail Stop 5, Huntsville, Alabama 35807.

therein; Crossman, *et al.*, in press;) provide, as a by-product, a convenient and relatively simple means of identifying redundant elements of the biota. At some stage in their computation, most agglomerative clustering methods involve the computation of a matrix of similarity or distance coefficients. This matrix shows pairwise similarities between all items to be clustered, in this case samples. Several such matrices may be computed, for example, one for each major group of organisms found in the study area. When two matrices contain the same number of elements, they may be compared by computing a product-moment correlation coefficient between their corresponding elements (Sokal and Rohlf, 1962; Kaesler, 1970). This method, adapted from numerical taxonomy where the coefficient is called the cophenetic correlation coefficient, was applied to matrices of Jaccard's coefficients (Jaccard, 1908) and is the basis of our method of estimating redundancy.

MATERIAL

We have chosen to study data from surveys of two streams. From 1956 until 1966, the Limnology Department of the Academy of Natural Sciences of Philadelphia made a series of twelve surveys of a portion of the upper Potomac River. The purpose of those surveys was to assess possible damage to the aquatic environment caused by the Dickerson Power Plant of the Potomac Electric Power Company. For this reason, the surveys were timed to coincide with the initial construction of the plant and with its enlargement by construction of additional units. Surveys were made both in June, when the water level in the river was high, and in August or September, when the water level was much lower.

Details of collecting, sampling analysis, and related information were given by Cairns (1966) and were summarized by Cairns and Kaesler (1969). The original surveys were based on the assumption that environmental stress or pollution would cause a reduction in number of species of aquatic organisms inhabiting the exposed area. Three stations were established: one upstream from the possible source of pollution, one just below the outfall of the power plant, and one about 10 km downstream where mixing was judged to be complete. No adverse effects of the thermal loading from the power plant were detected by either the method of Patrick (1949) or cluster analysis (Cairns and Kaesler, 1969; Roback, Cairns, and Kaesler, 1969; Cairns, Kaesler, and Patrick, 1970; Kaesler, Cairns, and Bates, 1971; Cairns and Kaesler, 1971; Kaesler and Cairns, 1972).

The Clinch River in southeastern Virginia has been the site of two recorded industrial spills (Cairns, *et al.*, 1971; Cairns, *et al.*, 1972; Crossman, *et al.*, in press).

The first major spill occurred when the dike surrounding a fly-ash holding-pond collapsed at Appalachian Power Company's power plant at Carbo, Virginia. Within less than an hour, $4.99 \times 10^5 \text{ m}^3$ of a slurry of $\text{Ca}(\text{OH})_2$ with a pH greater than 12 poured into Dump's Creek which joins the Clinch River only 0.8 km downstream. This caustic sludge equalled 40% of the daily flow of the Clinch River at the time and blocked the normal flow for several minutes. It also raised the water level several meters and forced some of the waste approximately 0.8 km upstream.

For 4.5 days following the spill, the alkaline sludge traveled downstream at a rate of approximately 1.5 km h^{-1} , killing essentially all the fish in its path. During this period, 162,000 sport and rough fish were killed in 39 km of the river in Tennessee before the polluted mass was diluted, dispersed, and neutralized by natural physical and chemical processes. The bottom fauna was also drastically reduced by the alkaline material (Cairns, *et al.*, 1971; Cairns, *et al.*, 1972; Crossman, *et al.*, in press).

RESULTS AND DISCUSSION

Correlation coefficients computed between corresponding elements of matrices of Jaccard's coefficient for all surveys of the upper Potomac River are shown in Table 1. The mean values at the bottom of the table have been corrected slightly from the values given by Kaesler and Cairns (1972).

The similarity matrix based on aquatic insects was found to have the highest average correlation with other similarity matrices, suggesting that of all the matrices it best summarized the results. Diatoms also gave a good representation of the total. Clearly one would not ordinarily plan to study both diatoms and other algae because of the high correlation between them, 0.688. Probably because of their mobility, fish were the least representative group of organisms studied. They have the advantage that they are ordinarily easily identifiable to the species level, whereas the ontogeny and details of the systematics of many larval aquatic insects remains to be studied. Based on these results, other invertebrates are probably the least desirable group for study because of the low average correlation coefficient they provide and because their taxonomic diversity, as a catch-all category, demands greater taxonomic expertise on the part of the investigators.

If we accept, at least tentatively, the advantages of working with aquatic insects, it is interesting to investigate the redundancy within data from this subset of the biota. We shall also compare again the aquatic insects with the other invertebrates.

Tables 2-4 give correlation coefficients between similarity matrices based on the total aquatic insect

Table 1. Correlation coefficients computed between corresponding elements of similarity matrices of Jaccard's coefficients from all surveys of the upper Potomac River; means of correlation coefficients for each group of organisms. (Modified from Kaesler and Cairns 1972)

	Diatoms	Other algae	Protozoa	Aquatic insects	Other invertebrates	Fish
Diatoms	—					
Other algae	0.688	—				
Protozoa	0.495	0.364	—			
Aquatic insects	0.603	0.556	0.581	—		
Other invertebrates	0.439	0.328	0.293	0.667	—	
Fish	0.332	0.459	0.145	0.425	0.386	—
Mean for group	0.511	0.479	0.376	0.566	0.423	0.349

fauna, the major orders of insects, other invertebrates, and gastropods from the three surveys of the Clinch River. For all 3 years, whether the stream was virtually unpolluted as in 1969, polluted by the spill of acid as in 1970, or devastated by flooding as in 1971, the Ephemeroptera had among the highest correlations with the total insect fauna, respectively 0.608, 0.566, and 0.785. In the data from the 1970 survey, Coleoptera had the highest correlation with the total insect fauna. This probably resulted from the comparative immunity of the beetles to the effects of the acid spill and the strong effect of the spill on the mayflies. During all surveys, however, the Coleoptera had high correlations.

Results of cluster analysis (Crossman, *et al.*, in press) and these high correlations suggest that the information provided by the Ephemeroptera alone would probably have sufficed for a study of stream recovery. Nevertheless, they did not represent perfectly the total insect fauna, and some other orders of insects may have provided useful information. The Coleoptera would be a good choice to study in addition to the Ephemeroptera because they were highly correlated with the distribution of the total insect fauna but were essentially uncorrelated with the Ephemeroptera during 1969 and 1970.

The matrices of Jaccard's coefficients from the combined Ephemeroptera and Coleoptera data from each of the surveys were compared with matrices computed from the total insect data. The correlation coefficients were 0.794, 0.817, and 0.860, strongly suggesting that data from these two orders of insects reveal most of the information that the total insect fauna carries.

During all surveys, the Gastropoda provided a good representation of the total other invertebrate fauna. In terms of new information provided, it is unlikely that the study of leeches, amphipods, decapods, bivalves, and others was worth the effort in this study. The correlation between similarity among stations based on other invertebrates and that among stations based on

aquatic insects was much lower than in the study of the upper Potomac River. It was not low enough to suggest a completely new dimension as was the case with the Coleoptera and Ephemeroptera.

The correlations between the Odonata and all other groups of organisms were very low and were sometimes slightly negative, but the correlation between the total insect fauna and the Odonata was also low so that the information they provided was of little additional value. During 1971 so few Odonata were found that it was impossible to compute a meaningful matrix of Jaccard's coefficients. Because of the high position of Odonata in the food chain, it is possible that information on their distribution might provide insight into the distribution of some fish, but this idea remains to be tested.

It appears that study of Ephemeroptera and Coleoptera would have provided nearly as much information about the recovery of the Clinch River as the study of the total benthic macroinvertebrate fauna. If so, the scope of the study could have been narrowed from 230 to 63 taxa with consequent savings in costs. Data on total other invertebrates and gastropods were almost totally redundant, so that study of 20 taxa of snails could have replaced the study of 52 taxa from many phyla. We would emphasize that these reductions in effort apply primarily to the use of cluster analysis in assessing stream recovery and not to measures of diversity. Furthermore, it is not yet clear to what extent these results can be generalized, and further testing using data from other environmental settings is indicated.

It follows from these results that judicious selection of the group of organisms to be studied can provide the investigators with different kinds of useful information. Study of Ephemeroptera and Trichoptera, both known for their proclivity to stream drift, can tell about the effect of pollutional stress on organisms that are likely to make a rapid recovery. Data on the snails can give a picture of the group that is slowest to recovery. Cole-

Table 2. Correlation coefficients computed between corresponding elements of similarity matrices of Jaccard's coefficients from data from Clinch River survey of 1969; means of correlation coefficients for each group or organisms

	Total insects	Ephemeroptera	Coleoptera	Odonata	Trichoptera	Diptera	Plecoptera	Other invertebrates	Gastropoda
Total insects	—								
Ephemeroptera	0.608	—							
Coleoptera	0.542	0.067	—						
Odonata	0.191	0.103	-0.008	—					
Trichoptera	0.520	0.225	0.123	-0.092	—				
Diptera	0.526	0.291	0.028	0.073	0.312	—			
Plecoptera	0.595	0.226	0.312	-0.135	0.387	0.244	—		
Other invertebrates	0.392	0.218	0.244	-0.023	0.303	0.201	0.270	—	
Gastropoda	0.388	0.221	0.245	-0.036	0.264	0.213	0.266	0.943	—
Mean for group	0.470	0.245	0.194	0.009	0.225	0.236	0.271	0.319	0.313

Table 3. Correlation coefficients computed between corresponding elements of similarity matrices of Jaccard's coefficients from data from Clinch River survey of 1970; means of correlation coefficients for each group of organisms

	Total insects	Ephemeroptera	Coleoptera	Odonata	Trichoptera	Diptera	Plecoptera	Other invertebrates	Gastropoda
Total insects	—								
Ephemeroptera	0.655	—							
Coleoptera	0.575	0.034	—						
Odonata	0.224	0.135	0.053	—					
Trichoptera	0.490	0.079	0.171	0.056	—				
Diptera	0.434	0.024	0.150	0.024	0.131	—			
Plecoptera	0.355	0.237	0.116	0.029	0.046	0.009	—		
Other invertebrates	0.339	0.349	0.240	0.129	0.144	0.030	0.227	—	
Gastropoda	0.394	0.399	0.238	0.122	0.103	-0.010	0.261	0.820	—
Mean for group	0.430	0.228	0.197	0.096	0.152	0.099	0.160	0.292	0.291

Table 4. Correlation coefficients computed between corresponding elements of similarity matrices of Jaccard's coefficients from data from Clinch River survey of 1971; means of correlation coefficients for each group of organisms. Cluster analysis not performed on data for Odonata

	Total insects	Ephemeroptera	Coleoptera	Trichoptera	Diptera	Plecoptera	Other invertebrates	Gastropoda
Total insects	—							
Ephemeroptera	0.785	—						
Coleoptera	0.594	0.351	—					
Trichoptera	0.538	0.332	0.067	—				
Diptera	0.391	0.207	0.071	0.082	—			
Plecoptera	0.566	0.397	0.197	0.230	0.124	—		
Other invertebrates	0.253	0.119	0.343	0.083	-0.070	0.204	—	
Gastropoda	0.300	0.184	0.378	0.072	-0.086	0.249	0.918	—
Mean for group	0.490	0.339	0.286	0.201	0.103	0.281	0.264	0.288

optera were largely unaffected by the spill of acid, and Odonata are high in the food chain and may yield information about some fish and other organisms at similar trophic levels.

Acknowledgements—Funds for surveys of the Potomac River were provided by the Potomac Electric Power Company and carried out while one of us (Cairns) was a member of the Limnology Department of the Academy of Natural Sciences of Philadelphia. We are grateful to Dr. Ruth Patrick, Chairperson of the Limnology Department, for many helpful suggestions and advice and to the staffs of the Potomac Electric Power Company and of Sheppard T. Powell and Associates for assistance and many courtesies during the field work on the Potomac River. The funds for the Clinch River Study were provided by the American Electric Power Company. Mr. T. A. Abolin, Mr. R. E. Sentes, and Mr. R. G. McComas, Appalachian Power Company, cooperated and assisted in various phases of the Clinch River Studies. We are indebted to Dr. Kenneth L. Dickson for help in locating stations and generally helping to initiate the Clinch River Studies.

The financial support for the cluster analyses was provided by Biomedical Sciences Support Grant 489-5706-6 from The University of Kansas and Office of Water Resources Research provided by Grant A-054-Kan through the Water Resources Institute of The University of Kansas.

REFERENCES

- Beck W. M. (1954) Studies in stream pollution biology. I. a simplified ecological classification of organisms. *Q. Jl Fla Acad. Sci.* **17**, 211–227.
- Beck W. M. (1955) Suggested method for reporting biotic data. *Sewage ind. Wastes* **27**, 1193–1197.
- Brown S.-D. (1969) Grouping plankton samples by numerical analysis. *Hydrobiologia* **33**, 289–301.
- Buzas M. (1972) Patterns of species diversity and their explanation. *Taxon* **21**, 275–286.
- Cairns J., Jr. (1966) The Protozoa of the Potomac River from Point of Rocks to Whites Ferry. *Notul. Nat. Academy of Natural Sciences of Philadelphia* **387**, 1–11.
- Cairns J., Jr., Albaugh D. W., Busey F. and Chanay M. D. (1968) The sequential comparison index—a simplified method for non-biologists to estimate relative differences in biological diversity in stream pollution studies. *J. Wat. Pollut. Control Fed.* **40**, 1607–1613.
- Cairns J., Jr., Crossman J. S. and Dickson K. L. (1972) The biological recovery of the Clinch River following a fly ash pond spill. *Proc. 25th Purdue Ind. Waste Conf., Purdue Univ. Engng Bull.* **137**, 182–192.
- Cairns J., Jr., Crossman J. S., Dickson K. L. and Herricks E. E. (1971) The recovery of damaged streams. *Ass. Southeast. Biol. Bull.* **18**, 72–106.
- Cairns J., Jr. and Dickson K. L. (1971) A simple method for the biological assessment of the effects of waste discharges on aquatic bottom-dwelling organisms. *J. Wat. Pollut. Control Fed.* **43**, 755–772.
- Cairns J., Jr., Dickson K. L., Sparks R. E. and Waller W. T. (1970) A preliminary report on rapid biological information systems for water pollution control. *J. Wat. Pollut. Control Fed.* **42**, 685–703.
- Cairns J., Jr. and Kaesler R. L. (1969) Cluster analysis of Potomac River survey stations based on protozoan presence-absence data. *Hydrobiologia* **34**, 414–432.
- Cairns J., Jr. and Kaesler R. L. (1971) Cluster analysis of fish in a portion of the upper Potomac River. *Trans. Am. Fish. Soc.* **100**, 750–756.
- Cairns J., Jr., Kaesler R. L. and Patrick Ruth (1970) Occurrence and distribution of diatoms and other algae in the upper Potomac River. *Notul. Nat. Academy of Natural Sciences of Philadelphia* **436**, 1–12.
- Crossman J. S., Cairns J., Jr. and Kaesler R. L. (in press) Aquatic invertebrate recovery in the Clinch River following hazardous spills and floods. *Water, Bulletin Virginia Water Resources Research Center.*
- Crossman, J. S., Cairns, J. Jr., and Kaesler, R. L. (in press) The use of cluster analysis in the assessment of spills of hazardous materials. *Am. Midl. Nat.*
- Fjerdingstad E. F. (1966) Some remarks on a saprobic system. In: *Biological Problems in Water Pollution*. Public Health Service Publication 999-WP-25, 232–235.
- Gaufin A. R. and Tarzwell C. M. (1952) Aquatic invertebrates as indicators of stream pollution. *Publ. Hlth Rep.* **67**(1), 57–64.
- Gaufin A. R. and Tarzwell C. M. (1956) Aquatic macroinvertebrates communities as indicators of organic pollution in Lytle Creek. *Sewage Ind. Wastes* **28**, 906–924.
- Hynes H. B. N. (1962) The significance of macroinvertebrates in the study of mild river pollution. In: *Biological Problems in Water Pollution*. Public Health Service Publication 999-WP-25, 235–240.
- Jaccard P. (1908) Nouvelles recherches sur la distribution florale. *Bull. Soc. Vaud. Sci. Nat.* **44**, 223–270.
- Kaesler R. L. (1966) Quantitative re-evaluation of ecology and distribution of Recent foraminifera and Ostracoda of Todos Santos Bay, Baja California, Mexico. *Univ. Kansas Paleontological Contributions*, paper **10**, 1–50.
- Kaesler R. L. (1970) The cophenetic correlation coefficient in paleoecology. *Bull. geol. Soc. Am.* **81**, 1261–1266.
- Kaesler R. L. and Cairns J., Jr. (1972) Cluster analysis of data from limnological surveys of the upper Potomac River. *Am. Midl. Nat.* **88**, 56–67.
- Kaesler R. L., Cairns J., Jr. and Bates J. M. (1971) Cluster analysis of non-insect macro-invertebrates of the upper Potomac River. *Hydrobiologia* **37**, 173–181.
- Kolkwitz R. and Marsson, M. (1908) Oekologie der pflanzlichen Saprobien. *Ber. dt. bot. Ges.* **26**(a), 509–519.
- Margalef D. R. (1958) Information theory in ecology. *Gen. Syst.* **3**, 36–71.
- Mount D. I. (1969) Developing thermal requirements for freshwater fishes. In: *Biological Aspects of Thermal Pollution*, Vanderbilt University Press, 140–147.
- Patrick R. (1949) A proposed biological measure of stream condition, based on a survey of the Conestoga Basin, Lancaster County, Pa. *Proc. Acad. Nat. Sci. Philad.* **101**, 277–341.
- Roback S. S., Cairns J., Jr. and Kaesler R. L. (1969) Cluster analysis of occurrence and distribution of insect species in a portion of the Potomac River. *Hydrobiologia* **34**, 484–502.
- Simmons G. M., Jr. (1972) A preliminary report on the use of the sequential comparison index to evaluate acid mine drainage on the macrobenthos in a preimpoundment basin. *Trans. Am. Fish. Soc.* **101**, 701–713.
- Sokal R. R. and Rohlf F. J. (1962) The comparison of dendrograms by objective methods. *Taxon* **11**, 33–40.
- Wilhm J. L. and Dorris T. C. (1968) Biological parameters for water quality criteria. *Bioscience* **18**, 477–481.
- Williams W. T. (1971) Principles of clustering. *Ann. Rev. Ecol. Systematics* **2**, 303–326.