Ecological Indices and Graphical Modeling of Factors Influencing Benthic Populations in Streams

Florentina Bunea Peter Guttorp Thomas Richardson

NRCSE

Technical Report Series

NRCSE-TRS No. 036

The **NRCSE** was established in 1996 through a cooperative agreement with the United States Environmental Protection Agency which provides the Center's primary funding.

EPA

Ecological Indices and Graphical Modeling of Factors Influencing Benthic Populations in Streams

Florentina Bunea, Peter Guttorp and Thomas Richardson

December 8, 1999

1 Introduction : The scientific problem

To assess damage to biological systems, it is important to describe in qualitative and quantitative terms how complex biological and ecological systems respond to specific activities of human society. A first step in doing so is to develop indices (metrics) that can capture the effect of human disturbance. At least three multimetric indices have been proposed for benthic invertebrates: the invertebrate community index (ICI: Ohio EPA 1988; Yoder and Rankin 1995 a, b); the rapid bioassessment protocol III (Plafkin et al. 1989) and the benthic index of biological integrity (B-IBI : Karr and Kerans 1992; Kerans and Karr 1994; Fore et al. 1996; Rossano 1996; Karr 1998).We are going to focus here on the metrics composing the B-IBI. For the Puget Sound study, Karr and Chu (1994) suggested the use of the following metrics :

- Metric 1 : Total number of taxa
- Metric 2 : Number of Ephemeroptera taxa
- Metric 3 : Number of Plecoptera taxa
- Metric 4 : Number of Trichoptera taxa
- Metric 5 : Number of long-lived taxa
- Metric 6 : Number of intolerant taxa
- Metric 7 : Percentage of individuals of tolerant taxa
- Metric 8 : Percentage of predator individuals
- Metric 9 : Number of clinger taxa
- Metric 10 : Percentage dominance of first three taxa

The B-IBI is simply the sum of all these metrics. However, since seven of them are discrete variables (counts) and three of them are continuous (percentages), the metrics are first transformed into variables of the same type, categorical in this case. Following Karr and Chau 1997, we consider three levels for each variable. The criterion according to which we find the splits in the range of each metric is therefore a first important factor to be taken into consideration when studying the variability of the B-IBI. We describe in Section 2 two statistical methods of finding the cut off points. We study the variability of the B-IBI based on each of these newly proposed methods and on a third method, suggested by Karr and Chu. We assume for this a statistical model for the data collected in the Puget Sound lowland area that takes into account the biological variability of the number of organisms at each sampled site. We compare the results based on this model with those based on the model suggested in Karr 1994 in which the number of individuals at each site is considered fixed. We formalize this in Section 2.

Also, we would like to understand the sensitivity of each metric to human influence. Human influence is mediated by a number of different processes. Rossano (1995) used a qualitative index which incorporated the following factors: amount of effluent present at a site; type of effluent (agricultural/domestic, raw sewage/industrial); proximity of dams, weirs, levees; type of riparian vegetation. Karr and Chau, 1997, used as a measure of human influence the percentage of impervious area, which is a weighted average of certain measures of urbanization (see, for example, May (1996)). These measures of human influence were used primarily to validate the IBI, by demonstrating that higher human influence was associated with lower IBI scores. For this purpose it is sufficient to have a reliable indicator of human interference. Given such validation, it is natural to consider whether certain types of human influence are more or less directly related to particular metrics, and further, whether certain forms of human behavior have differential effects on the component indices of the IBI. Partial answers to these questions may help in identifying important aspects of human influence, facilitating the development of better predictions of IBI score, and possibly providing insight into the nature of the mechanisms involved. Thus we will model the relationship between the metrics composing the IBI and the covariates measuring human influence and we present this in Section 3.

2 The Index of Benthic Biological Integrity : a Statistical Analysis

2.1 Statistical methodology

Our analysis is based on the 1994 data set containing 31 sampled sites in the Puget Sound lowland area. At each site, there are up to 81 possible taxa . The sampling protocol consisted in the collection of three samples at each site, taken under very similar conditions in three consecutive days. Thus, the whole data set can be regarded as a 81×93 table of counts. The B-IBI is simply a statistic computed on this data. The components of the variability of the B-IBI are best understood if we express it as the following functions' composition :

$$\text{data} \xrightarrow{f} (M_1, \ldots, M_{10}) \xrightarrow{g} (M_1', \ldots, M_{10}') \xrightarrow{h} \text{B-IBI}$$

where f is the function through which the ten metrics are computed, g is the function that discretizes the metrics, and h is just the arithmetic mean of the ten discretized metrics. Thus, with g defined by

$$M_i^{'} = \begin{cases} 1 & \text{if } M_i < c_i^1 \\ 3 & \text{if } c_i^1 \leq M_i < c_i^2 \\ 5 & \text{if } M_i \geq c_i^2 \end{cases}$$

it is clear that, in order to study the statistical properties of the B-IBI, assuming that f is given, both distributional assumptions on the data and knowledge of the cut off points are needed.

Karr and Chau 1997 suggested that the two dividers of the range of each metric should be taken where the biggest change in the metric, along the human influence gradient, occurs. This method, to which we shall refer in the sequel as the JK method, seems to be open to subjectivism, since two different researchers can have different opinions on what "the biggest change" means. We suggest a method that tries to objectivize the previous one . We considered the percent impervious area as a measure of the human influence, and we then used it as the response in a regression tree model having a metric as a covariate. We thus have 10 regression models, one for each of the ten metrics. The nodes of the tree can be then taken as cut off points. At each node , the squared error biased when predicting the response from the metric under consideration is minimized. We found that for this data set, the method we described above, and to which we will refer throughout this paper as the $CART$ method, can distinguish up to 5 classes per metric. However, the present analysis was carried out using 3 classes only (obtained from the best two nodes, where "best" has the sense described above). Nevertheless, either $CART$ or JK introduce an additional element of randomness in the distribution of the B-IBI, since c_i^1, c_i^2, $i \in \{1, \ldots, 10\}$, depend now on an external variable. It would be interesting to see how the influence on the IBI of the splitting points found by either of the above methods would compare to the influence of splitting points found by a method that would just use the data themselves. The simplest way of giving a criterion of the latter type is to take the 33% and 67 % quantiles, for each metric, as the delimiters of the three classes. This method will be called $Quant$. We compare the IBI variability in these three cases in the next section.

At a given site S the data has a structure as in Table 1, with S_1, S_2, S_3 denoting the three subsamples and with n_{ij} denoting the number of organisms from Taxa i found in subsample S_j. The bootstrap analysis suggested in Karr 1994 assumed that at each site the data come from a multinomial distribution with given n_{++} and with the vector of probabilities estimated from the sample proportions.

We call this $Model$ 1. We used $Model$ 1 for this data set and we found that, on the one hand, this assumption doesn't seem to be realistic at some sites, and, on the other hand, at sites where we don't have a strong reason to believe that it shouldn't hold, the IBI variability is very small, suggesting that it can provide a fine distinction of site conditions. However, if one allows now the observed number of individuals to be treated as an incidence of a random variable itself, the IBI appears to have a much larger variability. More specifically we assume that each of the three subsamples are realizations of a $Multinom_{81}(m, \underline{p})$, where we estimate $\underline{p}$ by $\hat{\underline{p}} = (n_{1+}/n_{++}, \ldots, n_{81+}/n_{++})$ and we assume that m has a negative binomial distribution with parameters estimated from n_{+1}, n_{+2}, n_{+3} using the method of moments. We call this

	S_1	S_2	S_3	Row Totals
Taxa 1	n_{11}	n_{12}	n_{13}	n_{1+}
Taxa 2	n_{21}	n_{22}	n_{23}	n_{2+}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Taxa 81	$n_{81,1}$	$n_{81,2}$	$n_{81,3}$	n_{81+}
Column totals	n_{+1}	n_{+2}	n_{+3}	n_{++}

Table 1: Data Structure at a Generic Site

Model 2. The next section contains a comparison between the use of the two models. Also, the impact of the three methods of finding cut off points on the IBI variability is studied in the context of *Model* 2.

2.2 Puget Sound Lowland Area : Comparative Study

We used the three cut off methods to compute first the IBI at all sampled sites. We found that the differences in IBI for *JK* versus *CART* or *Quant* range from 0 to 8, whereas for *CART* versus *Quant* , with one exception, the range was from 0 to 2. This result suggests that a splitting rule that does not introduce external variation would give the same IBI values as the one that does, so one would hope that by using the former one would eliminate a source of unnecessary variability in the IBI structure. Table 2 below contains the IBI values at a site and the 95% empirical confidence intervals (CI) and credible regions (CR) for some selected sites. The confidence intervals were based on *Model* 1 when we used the *JK* method, and on *Model* 2 for all three splitting rules. The confidence intervals for sites with IBI bigger than 30 have lengths ranging from 2 to 4 suggesting that, for this data set, a change of 2 in the IBI would be a statistically significant one.

Site	JK	CART	Quant	JK Model 1
BA1	38	46	44	38
	[36,40]	[42,46]	[42,46]	[36,40]
BB2	36	44	42	36
	[34,38]	[40,46]	[42,46]	[36,38]
SC1	22	26	28	22
	[16,29]	[18,30]	[20,30]	[28,28]
KE1	18	14	14	14
	[12,18]	[10,16]	[10,16]	[26,28]
TH4	12	12	12	12
	[10,14]	[10,14]	[10,14]	[22,24]

Table 2: IBI variability

However for IBI smaller than 30, the situation changes drastically : the CI simulated from *Model* 1, do not contain the site values, and even worst, the lower end of these intervals is

6 to 12 more units than the site value (See Table 2). This major overestimation suggests that *Model* 1 does not capture the natural variability at sites with relatively low and low IBI. We considered one of the simplest models that would do so, namely *Model* 2. We refer again to Table 2: overall, the IBI variability has increased, as expected. However, for sites with IBI higher than 30, the CR have lengths ranging from 2 to 6, which is comparable to the CI obtained with *Model* 1. Thus, for well preserved sites, the assumption on the variability of the number of organisms at a site does not play a big role, as expected from biological considerations. On the other hand, for sites with modest IBI , this assumption seems very important to us. From a biological point of view, one expects damaged sites to have a more fluctuant benthic structure, thus motivating our choice for Model 2. For sites with an IBI value between 18 and 30, simulations from the second model yielded CR that would certainly contain the observed IBI at a site (unlike the CI based on the first model), but their length would be quite high, from 8 to 12 (See Table 2). Then, for very low IBI, hence for seriously damaged sites, the IBI variability seems to decrease again, the length of the CR being between 4 and 6 (See Table 2).

So, to conclude this section, recall that the study of the IBI requires finding splitting rules for the metrics composing the index and also making modeling assumptions on the data generating mechanism. The analysis we presented so far offers a partial answer to the first question and suggests an approach to the second one. Thus, we have proposed two methods for finding cut-off points . The advantage of using them is that they can be automated and are therefore objective. We note that when compared with the JK method, both $CART$ and *Quant* produced statistically different IBI values per site, for sites with high IBI. Also, it is clear that sites that have been exposed to different degrees of human influence have different variability, hence a more realistic model for our data should take this into consideration. In the next section we propose a model for the ten metrics, jointly, that includes site info in its construction.

3 Graphical Modeling of Factors Influencing Benthic Populations in Streams

In this section we study the relationship between the metrics composing the IBI and some variables(the covariates) that have been used to measure the degree of human influence. Following May(1996), we consider as covariates the percentage of area covered by the subsequent land-use types:

- Y_1 = Forested
- Y_2 = Agricultural land / Parks / Golf Courses / Open space
- Y_3 = Low-density residential (Rural)
- Y_4 = Medium density residential (Suburban)
- Y_5 = High density residential (Urban)
- Y_6 = Commercial / Industrial / Malls/ Business Parks

Hence, the problem we would like to address is the one of modeling $f(M_1, \ldots, M_{10}|Y_1, \ldots, Y_6)$, where f denotes the density distribution of the ten metrics, given the six covariates. Note that we always have:

$$\begin{aligned} f(M_1, \ldots, M_{10}|Y_1, \ldots, Y_6) &= f(M_1|M_2, \ldots, M_{10}, Y_1, \ldots, Y_6) f(M_2|M_3, \ldots, M_{10}, Y_1, \ldots, Y_6) \\ & \quad f(M_9|M_{10}, Y_1, \ldots, Y_6) f(M_{10}|Y_1, \ldots, Y_6) \end{aligned} \tag{1}$$

so if we could model the individual *univariate* conditional distributions, then we would have a model for the joint conditional distribution of the metrics. Recall that M_7, M_8 and M_{10} are percentages, hence they can be treated as continuous random variables on $[0, 1]$ or, using the transformation $\log x/(1-x)$, as continuous variables on R, whereas the rest of the metrics are discrete (counts). We assume that the three continuous variables are gaussian and that the rest follow a Poisson distribution. The data consists of measurements on the metrics and covariates at 30 sites in the Puget Sound area. Since at each site we had 3 independent samples, we fitted our models based on 90 data points.

The modeling procedure consists, in principle, in the following steps:

1. For each univariate conditional distribution determine which of the variables we condition on do indeed play a role, for our data set. Then, estimate each univariate conditional distribution.

2. Estimate the conditional joint distribution of the metrics by multiplying the estimated univariate conditional distributions obtained at the previous step.

For the first step above we suggest the following : For $1 \leq i \leq 10$ fit the regression of M_i on the variables appearing in its conditioning set, as described by (1). That is, fit three regression models with gaussian errors (corresponding to M_7, M_8 and M_{10}), and seven regression models with Poisson errors corresponding to the rest of the metrics. Discard the non-significant explanatory variables, and then estimate the conditional distribution of M_7, M_8 and M_{10}, respectively, by a gaussian one with mean and variance estimates based on the fitted regression, and estimate the conditional distribution of each discrete metric by a Poisson distribution with mean estimated through the fitted regression.

Hence, we do obtain an estimate of (1), based on a certain factorization of the density. We refer to a particular factorization of the density as to a *model*. To each such factorization corresponds a graph, encoding the conditional dependence structure of our variables. We show how to construct a graph based on such factorization in the next section and we elaborate on the advantage of associating a graph to a statistical model. Because of this association we are going to refer to the models we consider as to *graphical models*.

However, we have 10! decompositions as in (1), corresponding to the 10! ways in which we can permute the ten metrics. For each of them we can apply the previous strategy, so we have in the end 10! models. So, ideally, one needs to fit 10! models and have a criterion that enables one to select "the best" of them. There are numerous criteria used for model selection and we employ here the BIC (Bayesian Information Criterion). This means that, given a set of models $\{P_i : 1 \leq i \leq 10!\}$, we choose the one for which

$$\log \prod_{k=1}^{n} f_{P_i}(m_{1,k}, \ldots, m_{10,k}|y_{1,k}, \ldots, y_{6,k}) - D_i \log n/2 \tag{2}$$

is maximum, where n is the number of observations (90 in our case), $m_{j,k}$ denotes the value of metric j, $1 \leq j \leq 10$, at site k, $y_{1,k}, \ldots, y_{6,k}$ are the values of the covariates at site k and D_i denotes the number of unknown parameters in model P_i.

The space of models is quite large, so one needs some form of automated search in order to narrow down the set of possible graphical models. (See Spirtes et al. 1993). Graph-based searches of this type have been applied to ecological problems by Shipley (1995, 1997). However, to the best of our knowledge, the existing computer packages deal either with continuous variables or with discrete variables, but not we both. So, for our study, we limited ourselves to fitting a small subset of the 10! models, corresponding to some permutations of the metrics that appear to have more biological relevance than others.

3.1 A graphical model

In this section we present the model we found as being "the best" from the set of models we fitted and a rule of drawing the graph associated to a particular model. Firstly, let us note that the covariates sum to one, so one needs to transform them in oder to obtain unique solutions for the parameter estimates. We have tried various transformations and we found that the one with the best explanatory ability was $\arcsin\sqrt{y}$. In what follows we still use Y to denote the covariates, but one should keep in mind that they have been transformed. We computed the BIC for 40 models, corresponding to various permutations of the metrics and various transformation of the covariates, and the average BIC score was 568.3 with a standard deviation of 13.8. The model we selected has a BIC score of 608.4, and it corresponds to the the $\arcsin\sqrt{y}$ transformation on the covariates. We note that the model corresponding to the same permutation of the metrics, but with no covariates, has a BIC score of 550.81, indicating that including the information on the covariates in modeling the distribution of the metrics might help understanding better the data generating mechanism. The joint conditional distribution corresponding to this model is :

$$\begin{aligned} f(M_1, \ldots, M_{10} | Y_1, \ldots, Y_6) \;=\; & f_1(M_{10}|M_8) f_2(M_9|M_1, M_6, Y_1, Y_3, Y_6) \\ & f_3(M_8|M_1, M_4, M_6) f_4(M_6|M_3, M_4) \\ & f_5(M_7|M_2) f_6(M_5) \\ & f_7(M_4|M_1, M_2, M_3, Y_6) f_8(M_3|M_1, Y_6) \\ & f_9(M_2|M_1, Y_1, Y_6) f(M_1|Y_1, Y_6) \end{aligned} \tag{3}$$

where, for example, f_1 is a gaussian density with mean depending on M_8, f_2 is a Poisson distribution whose mean depends on M_1, M_6, Y_1, Y_3, Y_6 (and similarly for the other metrics). Let us also note that throughout our analysis we considered the covariates as being fixed (that is, we did not treat them as random variables). The graph in Figure 1 in the Appendix corresponds to this decomposition. A graph corresponding to the decomposition (3) can be drawn as follows: Each variable $M_1, \ldots, M_{10}, Y_1, \ldots, Y_6$ represents a node in the graph. The metric to the left of the conditioning sign appearing in a univariate conditional distribution is connected to all the variables appearing to the right of the sign by directed edges pointing to the conditioned metric. The covariates Y's, since treated as fixed, will not be connected to one another.

The graph facilitates understanding the dependence structure between our variables. For example, we can read directly off the graph that, for example, M_6 is independent of Y_6 given M_4. For the general rules of reading conditional independencies when given a particular DAG (directed acyclic graph) we refer the reader to Whittaker (1989). This opens up the possibility of explaining which of the metrics are directly related to covariates explaining human influence, which of these covariates are possibly redundant in giving insight on the biological dynamics, etc.

Also, very importantly, having a model (and an estimate), for the conditional distribution of the ten metrics will allow us to run simulations, which provide a means of studying the distributional properties of the 10 metrics and of the IBI. This would provide an alternative to the bootstrap analysis of the IBI carried out by Fore et al. (1994) and to the one suggested by us in the previous section. This is the next step in our analysis and it will complement the work we have done previously on the study of the variability of the IBI.

References

Fore, L.S. et al. 1996 Assessing invertebrate responses to human activities : Evaluating alternative approaches. J.N.Am.Benthol.Soc.15 : 212-231

Gauch, H. G. 1982. Multivariate Analysis in Community Ecology. Cambridge University Press, Cambridge,UK.

Karr, J.R. and E.W.Chau 1997. Biological Monitoring and Assesssement : Using Multimetric Indexes Effectively. EPA 235-R97-001. UW, Seattle.

Karr, J.R., and B.L.Kerans. 1992. Components of biological integrity : Their definition and use in development of an invertebrate IBI. Pages 1-16 in T.P.Simon and W.S. Davis,eds. Environmental Indicators: Measurement and Assessment Endpoints. EPA 905/R-92/003.US EPA ,Chicago.

Kerans, B.L. and J.R.Karr 1994. A benthic index of biotic integrity (B-IBI) for rivers of the Tennessee Valley. Ecol.Appl.4 : 768-785.

Kleindl, W.J 1996. A benthic index of biotic integrity for Puget Sound lowland streams. MS Thesis , UW, Seattle.

Lauritzen, S.L. 1996. Graphical Models. Oxford Statistical Science Series, 17. OUP.

May,C.W. 1996 Assessement of cumulative effects of urbanization on small streams in the Puget Sound lowland ecoregion. Ph.D. Thesis , UW, Seattle Plafkin, J.L, M.T. Barbour, K.D. Porter, S.K. Gross, and R.M. Hughes. 1989 : Rapid bioassessment protocols for use in streams and rivers : Benthic macroinvertebrates and fish. EPA/440/4-89-001. Assessment and Water Protection Division, US EPA, Washington, D.C.

Rossano, E.M. 1996 Diagnosis of stream environments with index of biological integrity. Museum of Streams and Lakes, Sankaido Publishers, Tokyo.

Rossano, E.M. 1995 Development of an index of biological integrity for Japanese streams (IBI-J). MS Thesis UW, Seattle.

Shipley, B. 1997. Exploratory path analysis with applications in ecology and evolution. AM.-NAT. 1997 vol. 149, no. 6, pp. 1113-1138.

Shipley, B. 1995. Structured Interspecific Determinants of Specific Leaf Axes in 34 Species of Herbaceous Angiosperms. Functional ecology. Apr 1995. v. 9 (2) pp. 312-319.

Spirtes, P., C. Glymour, and R. Scheines 1993. Causation, Prediction and Search. Lecture Notes in Statistics, Springer-Verlag.

Tabachnick, B.G. et al 1989 Using Multivariate Statistics, 2nd ed., Harper Collins, New York

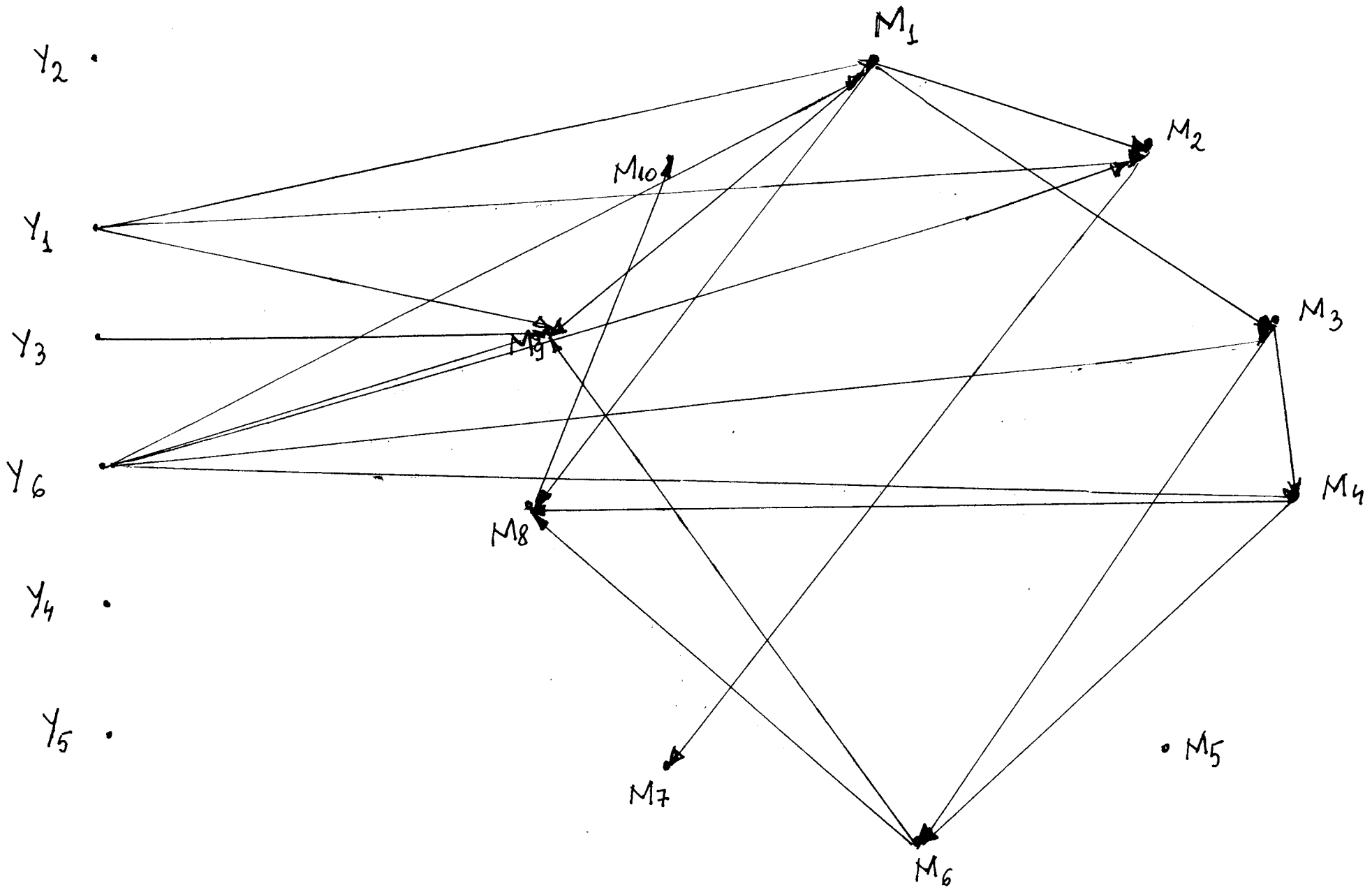
Ter Braak, C.J.F. 1986 Canonical correspondence analysis : A new eigenvector technique for multivariate direct gradient analysis. Ecology 67 : 1167-1179.

Whittaker, J. (1990). Graphical Models in Applied Multivariate Statistics, Wiley, NY.

Wright, S. (1934). The Method of Path Coefficients. Annals of Mathematical Statistics 5, 161-215.

Yoder, C.O., and E.T. Rankin 1995a. Biological criteria program development and implementation in Ohio. Pages 109-144 in W.S. Davis and T.P. Simon,eds. Biological Assessment and Criteria : Tools for Water Resource Planning and Decision Making. Lewis, Boca Raton, FL.

Yoder, C.O., and E.T. Rankin 1995b. Biological response signatures and the area of degradation value : New tools for interpreting multimetric data. Pages 263-286 in W.S. Davis and T.P. Simon, eds. Biological Assessment and Criteria : Tools for Water Resource Planning and Decision Making. Lewis, Boca Raton, FL.


M1
M2
M3
M4
M5
M6
M7
M8
M9
M10
Y1
Y2
Y3
Y4
Y5
Y6