4.3.1 Standardisations

The following standardisations were adopted for all invertebrate data analysis:

- For Trichoptera (cased caddis), which have a sliding-scale of types, the maximum number was set at 3- types, and if more than 3-types were recorded, they were included in the 3-type group for analysis.
- For cluster and ordination of invertebrate data:
  - If a taxon was present in < 5% of samples under investigation, it was excluded. Rare taxa contribute little to multivariate analysis and future model predictions are based on small sets of commonly occurring taxa (Marchant *et al*. 1997).
  - If < 6 taxa were present at a site the site was excluded.
  - If < 5 sites constituted a classification group, the group was excluded.

4.3.2 Statistical methods

Univariate and multivariate procedures were selected for analyses of biotic and abiotic data gathered in this study. Univariate procedures such as Analysis Of Variance (ANOVA) or the non-parametric equivalent (Kruskal-Wallis) test for significant differences between means or medians by comparing the variance between groups (Statistica). Such procedures are best applied to index or environmental data. Multivariate procedures are often used on biotic data, and consider each species/family to be a variable and the presence/absence or abundance of each species/family to be an attribute of a site or time (Norris & Georges 1993). Multivariate procedures include classification (clustering procedures), ordination and Discriminant Function Analysis (DFA). Classification techniques develop groups based on degree of similarity. Ordination reduces the dimensionality of a complex multivariate data.

4.3.2.1 Classification and ordination of invertebrate data

Sites were classified into Reference Groups based on their invertebrate communities. Invertebrate community data were transformed using the presence/absence transformation (PRIMER V.4) and the Bray-Curtis coefficient was used on these transformed data. Comparison of each sample with every other sample using this measure of similarity/dissimilarity leads to a triangular matrix, which can then be used in cluster and ordination analyses. Hierarchical agglomerative clustering, using group-average linking, was used on the data matrix. Ordination of faunal samples by multidimensional scaling (MDS) was undertaken and stress values used to assess the reliability of the MDS ordination. The distinguishing taxa responsible for the similarity within groups and the dissimilarity among groups were established using SIMPER (PRIMER V.4).

4.3.2.2 Environmental variables

The environmental variables distinguishing each group of reference sites were identified using Discriminant Function Analyses (DFA, Statistica). DFA facilitates the development of predictive models between groups of sites or taxa and environmental variables. It is assumed that the variables are independent and normally distributed. One of the major uses of DFA in the analysis of invertebrate studies is the

development of models that can be used to predict the membership of a site or group of taxa in a previously established classification of sites or taxa. A stepwise approach is recommended for finding the minimum subset of environmental variables that provides adequate prediction of group membership.

4.3.2.3 Univariate statistics

Two non-parametric analyses were undertaken on the SASS data since this was often not normally distributed. Kruskal-Wallis, a non-parametric equivalent of ANOVA, was used on environmental variables and index data [SASS4 Score, Number of Taxa and Average Score Per Taxon (ASPT]. The Kolmogorov-Smirnov test, a non-parametric alternative to the t-test, was used to test for differences between pairs of reference groups.

4.3.3 Invertebrate data: classification of reference sites based on composite reference community data

A composite reference community was derived by combining invertebrate data from each of three seasons, autumn, winter and spring and all three biotope-groups. Combining seasons for classifying sites has been adopted in both the United Kingdom and Australia and has been shown to be advantageous for site classification (Wright *et al.* 1993) and identification of outliers. Subsequent predictive models may be differentiated into seasons and/or biotopes, so that the expected fauna can be predicted for the appropriate single or paired season(s) or particular biotope.

The following steps were taken for analysis of invertebrate data:

1. Taxa recorded from all biotopes at a site during each of three sampling seasons, namely autumn, winter and spring, were combined into a composite reference invertebrate community.
2. The highest number of types were used for "type-dependent taxa" (Baetidae, Hydropsychidae and Trichoptera (cased-caddis), i.e. if 2 types of baetids were recorded in autumn, and 3 types of baetids were recorded in spring, then 3 types were used in the composite analysis.
3. Multivariate statistics (clustering and ordination, see section 4.3.2) were used to classify reference sites into groups (hereafter called "Reference Groups") based on the degree of similarity in their invertebrate communities. A Reference Group therefore represents a group of reference sites which have similar invertebrate communities.
4. Analyses were run on several *a priori* spatial groups (ecoregion and sub-region) and combinations thereof (Table 4.2), such that the homogeneity within groups could be tested, and the appropriateness of ecoregions and sub-regions for this geographic region (Mpumalanga) thereby tested. The various combinations included analyses:
   - within an ecoregion/sub-region
   - within an ecoregion (ignoring sub-region)
   - within a sub-region (ignoring ecoregion)
   - all combined (ignoring both ecoregion and sub-region)
5. Final Reference Groups were decided on examination of all the analyses.

The data was further examined to establish the:

- taxa responsible for the similarity within each Reference Group and the dissimilarity between Reference Groups.
- frequency of occurrence of each contributing taxon in each Reference Group.
- medians and ranges of SASS4 Scores, number of taxa and ASPTs for each Reference Group.
- environmental variables which provided maximum discrimination between Reference Groups (see section 4.3.4).

Table 4.2 Combinations of analyses undertaken for composite reference site data. H = Central Highlands, E = Great Escarpment Mountains, L = Lowveld. Sub-region: MS = Mountain Stream, FC = Foothill-cobble Bed, FG = Foothill-gravel Bed, RC = Rejuvenated Cascade, RF = Rejuvenated Foothill.

Ecoregion	**H**	**E**	**L**	**H+E**	**E+L**	**H+E+L**
Sub-region						
MS	✓	✓		✓		
FC	✓	✓			✓	✓
MS+FC	✓					
MS+FC+RC		✓		✓		✓
FC+FG+RC+RF			✓			
MS+FC+FG+RC+RF						✓

4.3.4 Environmental variables

Table 4.3 lists the environmental variables considered during this study and incorporated into the Discriminant Function Analysis (DFA) if shown to be potentially important discriminators during preliminary analyses. Variables were divided into four types:

- Catchment variables: including longitude, latitude, altitude, distance from source and stream order.
- Site variables: including channel pattern, stream width, habitat depths, geological and vegetation types and canopy cover.
- Habitat variables: including details on substratum richness, composition and dominance, the percentage of each substratum type, percentage embeddedness, the number and combination of biotopes, the percentage of each biotope present, and the percentage cover of algae and macrophytes.
- Water chemistry variables: including pH, temperature, conductivity, turbidity, dissolved oxygen and nutrients (total phosphorus, Kjeldahl nitrogen, nitrate+nitrite, ammonium and silica).

In cases where a value is given, it is taken as the mean value of the three sampling visits. Variables were tested for independence and, where necessary, data were log-transformed to approximate normality prior to analyses. Several variables were categorised into "types", the number of categories of which are given in parenthesis, and details of which are as per those in Table 4.1. or as below:

- Channel pattern (as per Table 4.1)

- Shallow-water habitat-type: three categories: 1: bedrock rapid, 2: bedrock rapid/cobble riffle mix and 3: cobble riffle
- Geological/lithostratigraphic type (as per Table 4.1)
- Vegetation type (as per Table 4.1)
- Canopy cover (as per Table 4.1)
- Substratum composition: four categories: 1: BR/B/CP/G, 2: BR/CP/G, 3: B/CP/G and 4: CP/G
- Substratum dominance (as per Table 4.1)
- Biotope combination: five categories: 1: all three biotopes, 2: SIC/SOOC + AQV/MV, 3: SIC/SOOC + G/S/M, 4: AQV/MV + G/S/M and 5: SIC/SOOC only.

Prior to DFA, variables were analysed using a non-parametric ANOVA (Kruskal-Wallis) using Reference Group membership as the factor variable (Statistica). In general, environmental variables that showed significant differences ($p < 0.05$) among Reference Groups were chosen for further analyses. DFA was run on two sets of Reference Groups, the first comprising three groups, and the second comprising five groups, including two sub-groups (see section 4.4.1). Whilst the development of predictive models was beyond the scope of this study, it was considered useful to identify those environmental variables which provide the greatest discrimination between Reference Groups. With the view to developing predictive models in the future, a subset of environmental variables which produced the lowest error in predicting Reference Group membership of a site in the DFA was identified (see section 4.4.2).

4.3.5 Separate- versus combined-biotope sampling

The primary objective of any biomonitoring programme is to establish the condition or health of a site or sites being assessed. To achieve this one needs a baseline with which data from a monitoring site may be compared. This interpretative tool is the function of reference conditions. Difficulties arise, however, when factors not related to anthropogenic modification and hence site disturbance, such as biotope availability and season, come into play. It is necessary to assess the extent to which the monitoring site is impacted, relative to the appropriate reference condition, in the absence of the influence of these factors.

Figure 4.4 illustrates diagramatically the interplay of these factors and indicates the options available for data analysis when these factors are taken into account. Option A is the combination used for deriving the composite reference condition (section 4.3.3) used for site classification. Option C is the combination for assessing the influence of biotope availability (section 4.3.5) and option B for assessing the influence of season (section 4.3.6).

Table 4.3 Environmental variables. Those that showed significant differences (KW: Kruskal-Wallis, $p < 0.05$, indicated with an ✓) among Reference Groups were chosen for Discriminant Function Analysis. If the Code is prefixed with a L, then the variable was $\log_{10}(x)$ transformed. A: 3 Reference Groups, B: 3 Reference Groups and 2 sub-groups.

Variable Type	Variable	Code	KW	
			A	B
Catchment variables	Longitude	LONG	✓	✓
	Latitude	LAT		✓
	Altitude	LALT	✓	✓
	Distance from source	LDIS	✓	✓
	Stream order (1-4)	ORD	✓	✓
Site variables	Channel pattern (1-3)	CHP	✓	✓
	Stream width	LW	✓	✓
	Shallow-water habitat: mean depth	LSAVG	✓	✓
	Shallow-water habitat-type (1-3)	SHType	✓	✓
	Deep-water habitat: mean depth[1]	LDAVG	✓	✓
	Geological/lithostratigraphic type (1-9)	GEOL		
	Vegetation type (1-9)	VEG	✓	✓
	Canopy cover (1-3)	CC		
Habitat variables	Substratum richness (1-4)	SUBRICH	✓	✓
	Substratum composition (1-4)	SUBCOMP	✓	✓
	Substratum dominance (1-9)	SUBDOM		✓
	% Bedrock	BR	✓	✓
	% Boulder	B	✓	✓
	% Cobble/pebble	CP	✓	✓
	% Gravel/sand/mud	GSM	✓	✓
	% Embeddedness	EMB		
	Biotope number (1-3)	BIOTNO		
	Biotope combination (1-5)	BIOTCOMB		
	% SIC/SOOC	SIC/SOOC	✓	✓
	% AQV/MV	AQV/MV	✓	✓
	% G/S/M	G/S/M	✓	✓
	% Algae	ALGAE	✓	
	% Macrophytes	MACRO		
Water chemistry variables	pH	PH	✓	✓
	Temperature (^oC)	TEMP	✓	✓
	Conductivity (mS m^{-1})	LCOND	✓	✓
	Turbidity (NTU)	LTURB		
	Dissolved oxygen (mg l^{-1})	LDO		
	Alkalinity (meq l^{-1})	LCACO3	✓	✓
	Total Phosphorus (mg P l^{-1})	LTP	✓	✓
	Kjeldahl nitrogen (mg N l^{-1})	LKN	✓	✓
	Nitrate+Nitrite (mg N l^{-1})	LNO3+NO2-N		
	Ammonium (mg N l^{-1})	LNH4-N		
	Silica (mg l^{-1})	LSI		