Chapter 3: Workpackage 1B: Development of a predictive bioassessment model for fish in rivers

3.1 Overview of chapter

Chapter 3 deals with the development of a predictive model to predict the composition of fish stocks in rivers. The reference condition approach was followed during this project.

3.2 Introduction

Three major approaches to the assessment of the ecological effects of pollution and landscape alteration on streams have been developed in the last two decades (Joy and Death, 2002). One approach is multimetric, where a number of individual indices are combined to measure biotic condition e.g. the Index of Biotic Integrity (IBI) (Karr, 1981; Gerritson, 1995), the ICI (Plafkin, 1989 for invertebrates-USEPA rapid bioassessment protocols) and the European Fish Index (EFI) (FAME CONSORTIUM, 2004). A second approach is predictive and compares fauna to those predicted by empirical models to occur in the absence of human impacts e.g. RIVPACS (Wright *et al*, 1984; Clarke *et al*, 1996; Norris 1996), AUSRIVAS (Parsons and Norris, 1996) and HABSCORE (Milner *et al*., 1995). The third approach is using artificial intelligence (AI) techniques, this is a relatively new approach to river quality monitoring based on attempts to model the expert knowledge of environmental conditions to predict the species that should occur using two complementary techniques, i.e. pattern recognition and plausible reasoning (Walley and Fontama, 2000).

3.2.1 Predictive modelling

Considerable effort has gone into developing predictive models of invertebrate assemblages for use in biomonitoring e.g. RIVPACS developed by the NRA in the UK (Wright, J.F. 1984 and 2000), in Canada, BEAST (benthic assessment of sediment) (Reynoldson *et al*, 1995) and in Australia, the Australian River Assessment System (AUSRIVAS) is the basis for the national River Health Program (Parsons and Norris, 1996). Data for an Irish version are currently being developed by the Limnology Unit Department, University College Dublin (Kelly-Quinn, *pers. comm*.). However, this effort in developing predictive models of invertebrate assemblages has not been matched in the field of predictive modelling of fish communities (Joy and Death, 2002). A regional predictive model of freshwater fish occurrence using 200 reference sites has been developed in New Zealand (Joy and Death, 2002). That model was successfully used to predict fauna expected at test sites using methods

based on the RIVPACS system and has been further refined using AI analysis (Joy and Death, 2003).

The predictive modelling process involves classifying the reference sites into groups based on the fish communities and then building a discriminant model to classify sites into these groups using the site environmental variables. This model is then used to associate test sites with suitable reference sites so that a comparison can be made between observed and expected communities. Ideally the variables used to associate the test sites with reference sites should not be influenced by human impact so that the predictions are the fish communities to be expected in the absence of impacts. However, the predictions are based on existing conditions so the expected communities are realistic. Once the scores are found for the reference sites they then give the basis for background or existing conditions and these become the criteria by which test sites can be assessed. The rest of the sites can then be run through the model and site-specific Observed (O)/Expected (E) scores calculated to enable comparison with biotic scores.

3.2.2 The Reference Condition approach

The Reference condition approach is the basis of the RIVPACS/ AUSRIVAS models. It is based on comparing a biological community found at a test site to the range of communities observed at a set of reference sites. The approach involves the selection of a large number of sites to represent the acceptable condition of the region and model building follows the process outlined below (Fig. 3.1).

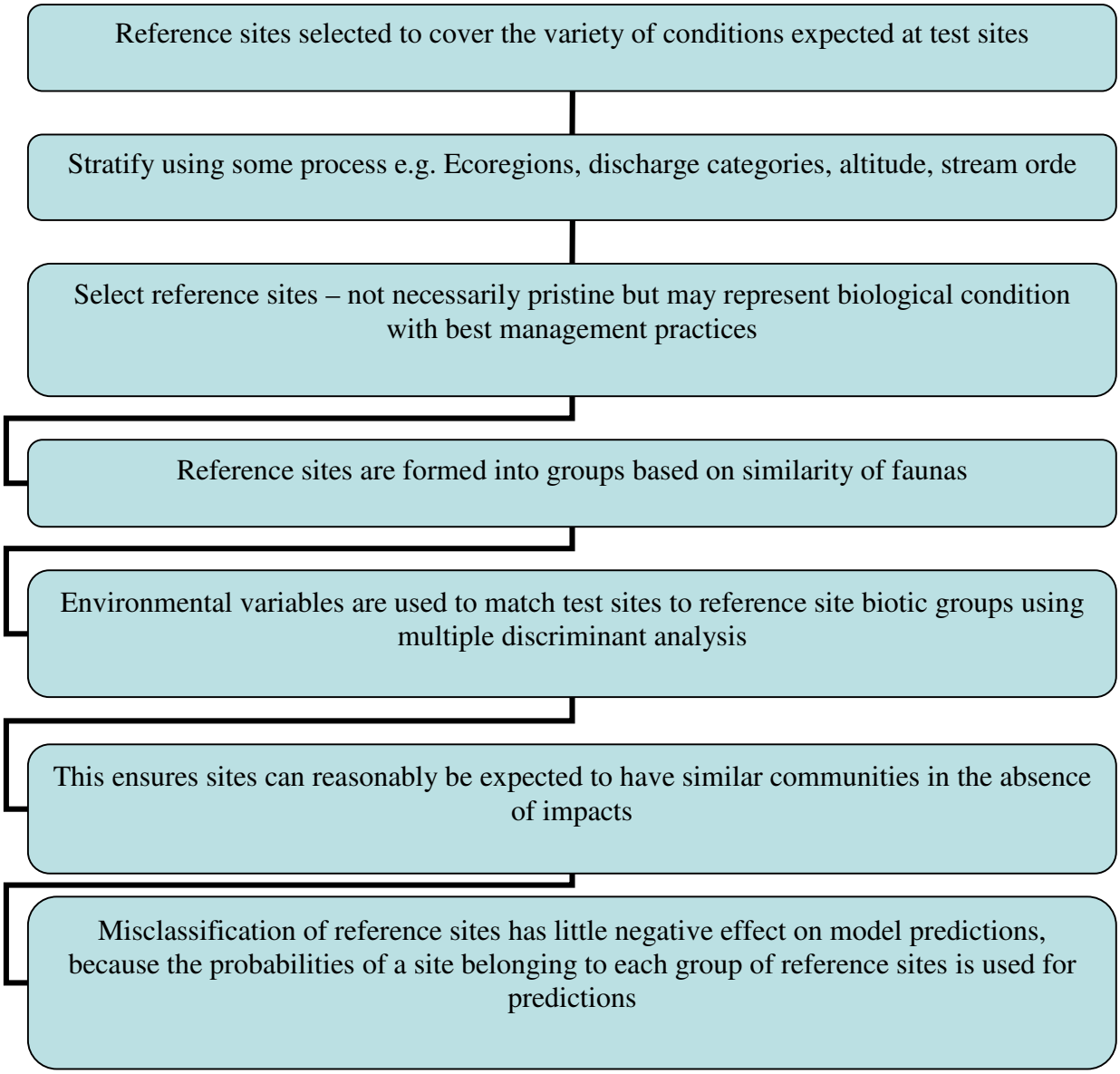

Fig. 3.1: Model building process

3.2.3 Aims of workpackage

The aim of this workpackage was to develop a predictive model, with known accuracy and precision, to predict the composition of fish communities in rivers (based on the physical and biotic elements of the aquatic ecosystem).

3.3 Methods

The statistical procedures/methods used in constructing RIVPACS and AUSRIVAS type predictive models using macroinvertebrates and fish have been described elsewhere (Wright *et al.*, 1984; Moss *et al.*, 1987; Wright *et al.*, 1993; Clarke *et al.*, 1996; Smith *et al,* 1999;

Simpson & Norris, 2000; Joy and Death, 2002) and the implementation described below followed similar steps using elements from both procedures.

3.3.1 Study area

High quality sites were selected using EPA quality ratings (Q-values). All sites achieving a Q-value of Q4-5 and Q5 were considered high quality or possible reference sites (Fig 3.2). The Q-values were validated by EPA biologists in conjunction with CFB staff.

3.3.2 Sampling regime

The standard methodology for the project includes fish stock assessment using electric fishing, kick sampling for macroinvertebrates, hydrochemical analysis and physical/habitat survey methodologies and has been described in the previous chapter. Surveys were carried out between July and September (to include capture of 0+ salmonids) when stream and river flows were moderate to low. Standard EPA methods were used to assess Q-values (Clabby *et al.*, 2000). A total of 118 of these high quality sites (Q4-5 and Q5) were used. Fish abundance data (number of fish m^{-2}) for these 118 sites was used to develop the model "with barriers" (this combines all high quality sites irrespective of whether or not a barrier or impediment to free passage existed downstream). Seventy eight sites were used to develop the model "without barriers" (i.e. only high quality sites where no impassable barriers to fish migration were located downstream were included). The barriers were identified using a GIS based data model for the quantification of the freshwater salmon habitat asset and for the determination of the quantity of habitat available to migratory salmonids (McGinnity *et al.*, 2003).

3.3.3 Statistical analysis-Predictive modelling

The high quality sites (Q5 and Q4-5) were divided into groups with similar fish communities using two-way indicator species analysis (TWINSPAN) (Hill, 1979) using PC-ORD (McCune and Mefford, 1997). This process classifies both samples and species simultaneously based on hierarchical divisions of reciprocal averaging ordination space. A number of studies have shown that this technique is useful as a preliminary analysis tool in freshwater biology both with stream invertebrates (Townsend *et* al., 1983; Ormerod & Edwards, 1987) and fish (Hayes *et al.*, 1989; Joy & Death, 2001; 2002 and 2003).

Discriminant analysis (SAS, 2000) was used to determine how well environmental variables account for the structure of biological groupings. Variables were entered as both log transformed and un-transformed data. The transformed variables were used if they improved the cross-validated classification rate (Clarke *et al.*, 1996). Cross-validation was used to check

whether sites were allocated to their correct groups. The cross-validation process (also known as jack-knife or leave-one-out validation) involves leaving out each site in turn, then re-building the model, and testing the withheld site to assess whether the site was predicted as belonging to the correct group. This jack-knife procedure has been shown to provide a robust and unbiased assessment when used with other similar models (Manel *et al.*, 1999; Manel *et al.* 2001; Olden *et al.*, 2002). A site was considered to be correctly classified if the probability of belonging to the correct group is higher than it is for the other groups. However, the actual value of this misclassification rate is not critical because all probabilities of group membership are used for predictions rather than just the group with the highest probability (Wright, 1995).

After optimisation of the discriminant functions, which allocate sites to the predetermined biological classification using their environmental characteristics (the classification is based on the assessment of the optimised suite of environmental characteristics), the next step was to predict the fish communities expected at a test site. To predict the assemblage expected at a site, the frequency with which individual taxa occur in each TWINSPAN group (i.e. the relative group frequency) was calculated as the number of sites where that taxon occurs divided by the total number of sites in the group. This is referred to as the probability of finding that taxon in that group. The overall probability of finding a taxon at a site is the relative group frequency weighted by the probability of membership in each of the three groups (Table 3.1).

Table 3.1: Example of RIVPACS type prediction of assemblage at a site. In this example, the species would be predicted to be present in the site based on a decision threshold of 0.5.

Twinspan group	Prob. site group membership	Prevalence in group	Prob. of species occurrence in group
A	0.7	0.8	0.56
B	0.2	0.5	0.10
C	0.1	0.2	0.02
Combined probability that species will occur at site			**0.68**

The final step in site assessment is to compare observed and expected faunas. The predicted fauna was compared with the observed taxa list following the procedure originally described by Wright *et al.* (1984). The probabilities of the predicted taxa were summed to give the

‘expected number of taxa’ (E). The number of species actually captured at a site, providing they were predicted to occur is the ‘observed number of taxa’ (O). The ratio of the observed to the expected number of taxa (*O/E*) and taxonomic composition is the output from the model (Moss *et al.*, 1987).

The number of taxa observed at high quality sites were compared with model predictions generating a distribution of reference site *O/E* ratios. Low *O/E* ratios are used to indicate sites under stress, while high ratios indicate sites with more species than expected, which may indicate sites of high conservation value (Wright, 1995). Determination of whether a site is impacted is judged based on the site’s *O/E* ratio compared with the distribution of *O/E* ratios for the reference sites.

To test for differences between the environmental site descriptors of the three TWINSPAN groups, the multi response permutation procedure (MRPP) (McCune & Mefford, 1997) using Euclidean distance measures was used. MRPP is a non-parametric procedure for testing the hypothesis that there are no significant differences between two or more groups and has the advantage over ANOVA that it is not reliant on multivariate normality and/or homogeneity of variances.

To validate the output from the model *O/E* ratios were calculated for sites not used in the model construction (i.e. the non-reference sites) and these values were then compared with Q-scores for the sites. The differences in *O/E* scores in the Q-score groups were examined using the Kruskal-Wallis test (NPAR1WAY procedure of (SAS, 2000)) and medians were compared using Tukey’s multiple range tests with the significance level set at 5%.

3.4 Results - Model development of reference sites with and without barriers, to fish migration, present downstream

3.4.1 Fish assemblages

Removal of rare species resulted in 11 taxonomic units used for model construction (Table 3.2). From this data set, the TWINSPAN analysis was taken to two levels and then two groups were combined at the second level as group sizes were too small and this resulted in three groups, (after validation using MRPP): Group 1 sites contained no salmon but trout occurred at all 47 sites. There were 56 sites in group 2 and trout and salmon occurred at all of the sites, however, stoneloach were absent from all sites and 3 spine sticklebacks were rare, occurring at only 5% of the sites. Group 3 contained only 11 sites, and stoneloach, trout and

salmon occurred at all sites in this group. Three spine sticklebacks and lamprey were more common in this group than the other two groups.

Table 3.2: Percentage of each taxon in each of the three groups

Species	Group 1	Group 2	Group 3
Eel	30	59	55
Minnow	19	9	9
3 spine	11	5	64
Stoneloach	26		100
Lamprey	11	4	45
Trout	100	100	100
0+ trout	100	91	64
older trout	94	96	82
Salmon		100	100
0+ salmon		89	64
older salmon		89	100

3.4.2 Relationships between fish assemblages and physical/chemical data

The TWINSPAN groups were then applied to the environmental variables and discriminant functions used to see if the biological groupings were separable by the site descriptors. Using discriminant analysis, 87 (74%) of the 114 reference sites were assigned correctly to their predetermined biological groups using cross-validation (Table 3.3). The highest error rate for the cross-validated classification rate was for the smallest group (3) with only 5 (46%) of the sites correctly classified.

Table 3.3: The cross-validated number and percentage of sites correctly classified into each of the three TWINSPAN groups by linear discriminant analysis using the concurrently measured environmental variables listed in Table 3.4.

Group (from group)	Predicted group membership (to group)			% of sites correctly predicted
	1	2	3	
1	**36**	8	3	76%
2	5	**46**	5	82%
3	0	6	**5**	46%

The environmental variables associated with the sites from the TWINSPAN analyses are summarized in Table 3.4 and Appendix 6. The MRPP analysis revealed significant differences between these groups based on the environmental site descriptors ($T = -6.91$, $P < 0.0001$).

Correlations between the discriminant factors and the environmental variables revealed the relationships between the environmental variables and the three groups (Table 3.4 and Fig 3.3: also see Appendix 7 for more details). Canonical axis 1 which accounted for 74% of the variation was strongly influenced by northing and barrier downstream, and negatively by stream order, sand, stream width, and geology (Table 3.4). Canonical axis 2, which accounted for a further 26% of the variation, was correlated with distance to tidal limit, the proportions of mud and silt, glide, total conductivity, water hardness and alkalinity (Appendix 8).

Table 3.4: Coefficients for the correlation between environmental variables with the first two axes of a canonical discriminant analysis. * *P*<0.05, ** *P*<0.01, ****P*<0.0001

Environmental variable	Correlations			
	Can 1 (74%)		Can 2 (26%)	
Easting	0.189	*	-0.095	
Northing	0.327	***	0.081	
Hydrometric area	-0.014		0.146	
Catchment area	-0.181		0.054	
Distance to source	-0.084		-0.173	
Distance to tidal	0.011		-0.397	***
Stream order	-0.308	***	0.098	
Altitude	0.218	*	-0.047	
Barrier downstream	0.816	***	0.088	
Bedrock	-0.023		0.181	
Boulder	-0.065		0.368	
Cobble	0.045		-0.174	
Gravel	0.167		-0.111	
Sand	-0.257	**	-0.165	
Mudsilt	0.026		-0.301	**
Riffle	0.037		0.229	*
Glide	0.023		-0.265	**
Pool	-0.060		-0.009	
Instream cover	0.195	*	-0.144	
Shade	0.040		0.284	**
Mean wetted width	-0.269	**	0.052	
Length of site	0.171		-0.003	
Maximum depth	-0.099		0.089	
Surface area	-0.021		0.033	
Geology	-0.253	**	-0.121	
Alkalinity	-0.021		-0.297	**
Total hardness	-0.054		-0.406	***
Conductivity	-0.026		-0.314	***
Landuse	-0.001		0.123	
Water levels	0.067		-0.058	
Velocity rating	-0.176		-0.232	*
Bank height lhs	0.088		0.004	
Fencing lhs	0.163		-0.299	**
Trampling lhs	-0.050		0.055	
Bankerosion lhs	0.101		-0.154	
Bank slippage lhs	0.233	*	-0.095	

Bank height rhs	0.098		0.057	
Fencing rhs	0.240	*	-0.490	***
Trampling rhs	-0.173		0.086	
Bank erosion rhs	0.125		-0.151	
Bank slippage rhs	0.155	*	-0.110	

3.4.3 Calculation of O/E ratios

The *O/E* ratios of the predicted and observed fish faunas were calculated using all taxa with probabilities >0.5. The expected number of taxa was obtained by summing the probability values for each of the taxa from the ranked list of weighted probabilities. This process was repeated for all reference sites, to give the distribution of *O/E* ratios (Fig. 3.4).

3.4.4 Model evaluation

To evaluate the ability of the model to assess the condition of a site, the non-high quality sites (Q-values ranging from bad (Q1) to good (Q4) quality) were run through the model and *O/E* ratios calculated in the same way as the reference sites and these values were compared. The mean *O/E*-value for all reference sites was close to unity at 0.89 (SE 0.018) which suggested that the model produced unbiased estimates of the number of taxa expected to occur at a site. The mean *O/E* ratio for the test sites was however, significantly lower at 0.66 (SE 0.01) (ANOVA $F_{1,467} = 95.6$, $P < 0.0001$).

Next the *O/E* values were grouped by Q-values where these were available to see if there was a relationship between *O/E* score and Q-score. The *O/E* scores were significantly different for the 6 Q-score groups ($F_{5,464} = 17.84$, $P < 0.0001$) the Tukey's mean test revealed that the differences were between the lower Q-scores but some showed overlap occurred at the higher scores (Fig. 3.5).

3.4.5 Probability cut-off levels for inclusion of taxa

The use of the 0.5 probability level followed the AUSRIVAS protocol rather than the more stringent 0.0 (i.e., all probabilities) level used in Britain with the RIVPACS models. Hawkins *et al*. (2000) found that when these two cut-off levels were compared, the 0.5 cut-off level yielded more robust model outputs. The 0.5 threshold was used in these models.

3.5 Results - model development of reference sites without barriers

Reference quality sites above artificial barriers presented a problem as they could not be considered "reference" with the barrier impact. Therefore a new predictive model was constructed without sites above barriers. The model construction and validation followed the process in the first model described in the methods section. There were 78 sites with scores of

Q4-5 and Q5 without barriers and these were run through a TWINSPAN analysis and four groups were identified (Table 3.5). Group 1 sites contained only eels, minnows and trout. Group 2 sites contained eels, minnows, sticklebacks, trout and salmon. Groups 3 sites contained all species and group 4 sites contained all species except salmon.

Table 3.5: Percentage of each taxon in each of the four groups

Group	1	2	3	4
Eel	40	60	70	08
Minnow	30	100	10	20
3 spine stickleback		100	70	80
Stoneloach			90	20
Lamprey (juveniles)			60	40
Trout (total)	100	100	100	100
0+ trout	100	90	50	100
1+ & older trout	100	90	90	80
Salmon (total)		100	100	
0+ salmon		90	50	
1+ & older salmon		90	100	

The discriminant model, which excludes sites affected by barriers to fish migration showed good discrimination (overall cross-validated error rate 28%) see Table 3.6. The variables decided on after using the set of variables shown in Table 3.7.

Table 3.6: The cross-validated number and percentage of sites correctly classified into each of the four TWINSPAN groups by linear discriminant analysis using the concurrently measured environmental variables listed in Table 3.7.

	Predicted group membership (to group)				
Group (from group)	**1**	**2**	**3**	**4**	**% of sites correctly predicted**
1	**7**	1			87.5 %
2	2	**48**	1	4	87.2 %
3	0	3	**7**		70 %
4		2	1	**2**	40 %

Table 3.7: Variables selected for use in the second model after stepwise discriminant analysis.

Variable	Partial R^2	F Value	Pr > F	Wilks' Lambda	Pr < Lambda
Streamorder	0.29	9.96	<.0001	0.71	<.0001
Tothardness	0.26	8.43	<.0001	0.53	<.0001
Easting (longitude)	0.25	7.85	0.0001	0.40	<.0001
Fencingrhs	0.24	7.36	0.0002	0.30	<.0001
Shade	0.15	4.21	0.0085	0.26	<.0001

Sand	0.13	3.57	0.0183	0.22	<.0001
Disttidal	0.14	3.57	0.0184	0.19	<.0001
Conductivity	0.18	4.86	0.0041	0.16	<.0001
Altitude	0.16	4.23	0.0085	0.13	<.0001
Maxdepth	0.13	3.14	0.0312	0.12	<.0001
Riffle	0.15	3.77	0.0148	0.10	<.0001
Gravel	0.16	4.08	0.0103	0.08	<.0001
Bankheightrhs	0.12	2.76	0.0496	0.07	<.0001
Northing	0.11	2.64	0.0575	0.06	<.0001
Waterlevels	0.16	3.73	0.0158	0.05	<.0001
Bankerosionrh	0.12	2.61	0.0598	0.05	<.0001
Bankerosionlh	0.12	2.68	0.0553	0.04	<.0001
Landuse	0.11	2.31	0.0857	0.04	<.0001
Fencinglhs	0.11	2.31	0.0858	0.03	<.0001

All 470 sites including those with barriers were run through the model and the O/E ratios were calculated and are shown in Figure 3.6. The mean O/E ratios for the sites without barriers (256 sites) are shown in Figure 3.7.

3.6 Discussion

The "with and without barriers" discriminant model assessment showed the percentage of sites correctly assigned to bio-groups and the distribution of reference site O/E scores was similar to many published RIVPACs and AUSRIVAS models produced in the UK, Australia and New Zealand using fish and invertebrates (Smith *et al.*, 1999; Wright *et al.*, 2000, Joy & Death, 2002; Joy & Death, 2003). This suggests the model produced here is robust and up to the standard of other similar models in use worldwide.

The correlation with Q-ratings was strong and positive but there was not a significant difference between the high quality sites and the "moderate" quality (Q3-4) sites. There are two probable interrelated explanations for this:

1). The Q-ratings are calculated using invertebrate communities which are mainly influenced by local conditions and the site watershed, on the other hand the fish are influenced by conditions upstream and downstream of the site.

2). The variable which had the strongest influence on the fish communities was the presence of a barrier to fish migration, this means that a number of the high quality sites may be influenced by restricted fish passage possibly negating their reference site status where the barrier is man-made (if the barrier is natural it doesn't affect a site's reference status). Both these explanations are liable to be influencing the results by lowering the high quality site scores and raising the test sites scores thus causing the lack of discrimination between high and test score.

The model constructed with high quality sites without barriers showed a clear separation between the four TWINSPAN groups although with only 78 sites some groups were small. The distribution of mean O/E values for all the sites after being run through the new model showed no better discrimination between Q-classes than the original model. The reason for the lack of discrimination between Q-classes by the predictive fish model is likely to be related to the Q-values being calculated using invertebrates.

As noted above the assessment of sites using invertebrates and fish are likely to be different due to differences in the scale of environmental influences especially because invertebrates are influenced primarily at the proximal or reach scale and secondly by catchment process upstream of the site, while fish communities are the product of these influences but also those downstream as some migratory species come form the sea. Thus, the fish communities with anadromous and catadromous species are indicators of the whole river from source to the sea, whereas the invertebrates integrate proximate and upstream influences. Given these differences it is not surprising that the two assessments do not totally agree.

The use of non-biological criteria for defining reference sites would largely get around the problems outlined above and then fish communities will allow for site assessment at the whole river scale. Furthermore, because only species presence/absence data were used influences on fish abundance would not be picked up in this process.

3.7 Recommendations

- Expansion of the data set to allow for segregation and classification of sites upstream and downstream of natural barriers to fish passage.
- Identification of high quality reference sites based solely on fish.