Hydrobiologia (2006) 572:59–70 © Springer 2006
R.H. Norris, R. Marchant and A. Milligan (eds), Evaluation and Application of Methods for Biological Assessment of Streams
DOI 10.1007/s10750-006-0252-7

Local stream habitat variables predicted from catchment scale characteristics are useful for predicting fish distribution

James Mugodo[1], Mark Kennard[2,*], Peter Liston[1], Sue Nichols[1], Simon Linke[1], Richard H. Norris[1] & Mark Lintermans[3]

[1]*Cooperative Research Centre for Freshwater Ecology, University of Canberra, Canberra, 2601, Australia*
[2]*Cooperative Research Centre for Freshwater Ecology, Centre for Riverine Landscapes, Faculty of Environmental Sciences, Griffith University, Nathan, Queensland, 4111, Australia*
[3]*Wildlife Research & Monitoring, Environment ACT, Canberra, 2602, Australia*
*(*Author for correspondence: E-mail: m.kennard@griffith.edu.au)*

Received 22 September 2004; in revised form 9 May 2006; accepted 21 May 2006; published online 23 August 2006

Abstract

South-east Queensland (Australia) streams were described by 21 local habitat variables that were chosen because of their potential association with fish distribution. An Assessment by a Nearest Neighbour Analysis (ANNA) model used large-scale variables that are robust to human influence to predict what the values of each of the 21 local habitat variables at each site would be without modification from human activity. The ANNA model used elevation, stream order, distance from source and longitude to predict the local habitat variables; other candidate predictor variables (mean rainfall, latitude and catchment area) were not found to be useful. The ANNA model was able to predict five of the 21 local habitat variables (average width, sand (%), cobble (%), rocks (%) and large woody debris) with an R^2 of at least 0.2. The observed values of these five local habitat variables were used to model the distributions of individual fish species. The species distribution models were developed using logistic regression based on a subset of the data (some of the data were withheld for model validation) and a forward stepwise model selection procedure. There was no difference in predictive performance of fish distribution models for model predictions based on observed values and model predictions based on ANNA predicted values of local habitat variables in the withheld data (p-value = 0.85). Therefore, it is possible to predict the suitability of sites as habitat for given fish species using estimated (estimates based on large-scale variables) natural values of local habitat variables.

Introduction

The physical habitat of many rivers worldwide has been degraded by human activities (Gorman & Karr, 1978; Imhof et al., 1996; Kauffman et al., 1997; Harper & Everard, 1998; Maddock, 1999; Gafny et al., 2000; Hall et al., 2002). The physical habitat, defined as the living space of instream biota, is spatially and temporally dynamic due to the interaction of the structural features of the channel (channel size, channel shape, gradient, bank structure and substrate) and the hydrological regime (Maddock, 1999). The state of this living space will influence biotic structure and organization within rivers (Richards et al., 1996; Kauffman et al., 1997; Richards et al., 1997; Maddock, 1999; Davies et al., 2000). The state of the habitat is influenced by factors operating at several spatial and temporal scales (Frissell et al., 1986; Imhof et al., 1996; Richards et al., 1996; Davies et al., 2000). At the catchment scale, geology and climate influence the habitat at the reach scale by affecting stream hydrology, sedimentation, nutrient inputs and channel morphology (Schumm & Lichty, 1965; Knighton, 1984; De Boer, 1992; Richards et al., 1996). In addition, operating at the local scale, human influences, particularly land use and land management practices, also influence reach scale habitat (Richards et al., 1996). Therefore, in the absence of human impacts it is conjectured that the physical habitat of a site will largely be predicted by catchment scale characteristics.

Physical habitat data in the absence of human disturbance are required for the assessment of habitat condition at a site (Maddock, 1999; Davies et al., 2000). Information on unimpacted habitat state would also be useful for predicting the potential distribution of different fish species based on their habitat requirements (Imhof et al., 1996). Furthermore, the observed and potential (predicted) habitats may be compared with the habitat requirements of a given species to ascertain whether the needs of the species are satisfied under the observed site conditions and whether they can be met if the site habitat is restored to its potential condition (Imhof et al., 1996; Maddock, 1999). However, pre-disturbance local habitat data are usually unavailable and have to be predicted.

Most river-based predictive studies of species distribution have been limited to predicting species distribution based on the current habitat variables (Armitage et al., 1987; Richards et al., 1996; 1997; Olden & Jackson, 2002; Bond & Lake, 2003). Thus, the usefulness of predicted local habitat variables as input for species distribution models is largely untested in rivers. Continuous variables are generally more desirable for fitting models than categorical variables because the latter require a larger number of parameter estimates (i.e. a parameter has to estimated for each category in the variable), increasing the chances of less stable models especially where the model is fitted to a small dataset. Previous studies have predicted the occurrence of habitat categories (Jeffers, 1998; Davies et al., 2000) but our study attempts to predict the values of continuous local habitat variables for use in fish distribution models. The performance of fish models based on predicted local habitat variables may be reduced because the species distribution predictions will be affected by errors in the species distribution model as well as errors in the predicted local variables. If model performance on predicted local variables is as good as the performance based on actual local habitat variables then the former are useful for predicting fish distribution.

In addition to the influence of local habitat characteristics, fish distributions may also be influenced by environmental factors at other spatial scales. Riverine habitats are arranged in a strongly hierarchical manner (Frissell et al., 1986) and thus, various habitat elements at one scale may act as determinants of species composition or abundance at other subordinate scales. Barriers to movement (e.g. cascades and waterfalls) or connectivity between larval and adult habitats (e.g. relative proximity to estuarine spawning areas for catadromous species) are examples of landscape filters (sensu Poff, 1997) that may determine which subset of the total species pool potentially occur at a local scale. Variation in local habitat characteristics may then determine which of this subset of species may actually be present. Fish distributions are thus constrained by large-scale and local-scale environmental factors and interactions with the biology (e.g. resource requirements, life cycle and movement) of the available species pool (Poff & Ward, 1990; Poff & Allan, 1995; Schlosser, 1995; Schlosser & Angermeier, 1995).

In this study, logistic regression models (McCullagh & Nelder, 1989) of fish species' presence–absence distribution are developed using those observed local habitat variables that can be accurately predicted from catchment-scale variables using the ANNA model (Linke et al., 2005). The fish models are then used to make predictions on samples that were withheld from the model development stage. The predictions are based on both observed local habitat variables and predicted local habitat variables of the withheld samples to determine whether the predictive performances are similar. In this study, local habitat or site habitat refers to the reach-scale habitat (~10 × wetted width).

Methods

Study area and site selection

A full description of the study area, criteria for site selection and methods for sampling fish and estimating environmental variables is available in Kennard et al. (2006) and Pusey et al. (2004) and only a brief overview is given here. The study area was confined to three river basins in coastal south-east Queensland, Australia (Fig. 1). Relatively few rivers of south-east Queensland are free of human disturbance. Study sites were therefore selected to represent the best condition available (i.e. least disturbed) within the study rivers, while also ensuring that such sites were arrayed sufficiently widely throughout each catchment to encompass as much of the biological and environmental variation as possible. Sampling was also limited to river reaches that were wadable and hence could be sampled effectively by electrofishing (i.e. generally less than 1.5 m maximum depth). Sampling was conducted during periods at or near base-flow conditions to avoid extreme flow events causing bias in our fish sampling or measurement of local habitat variables. Data used in this analysis were

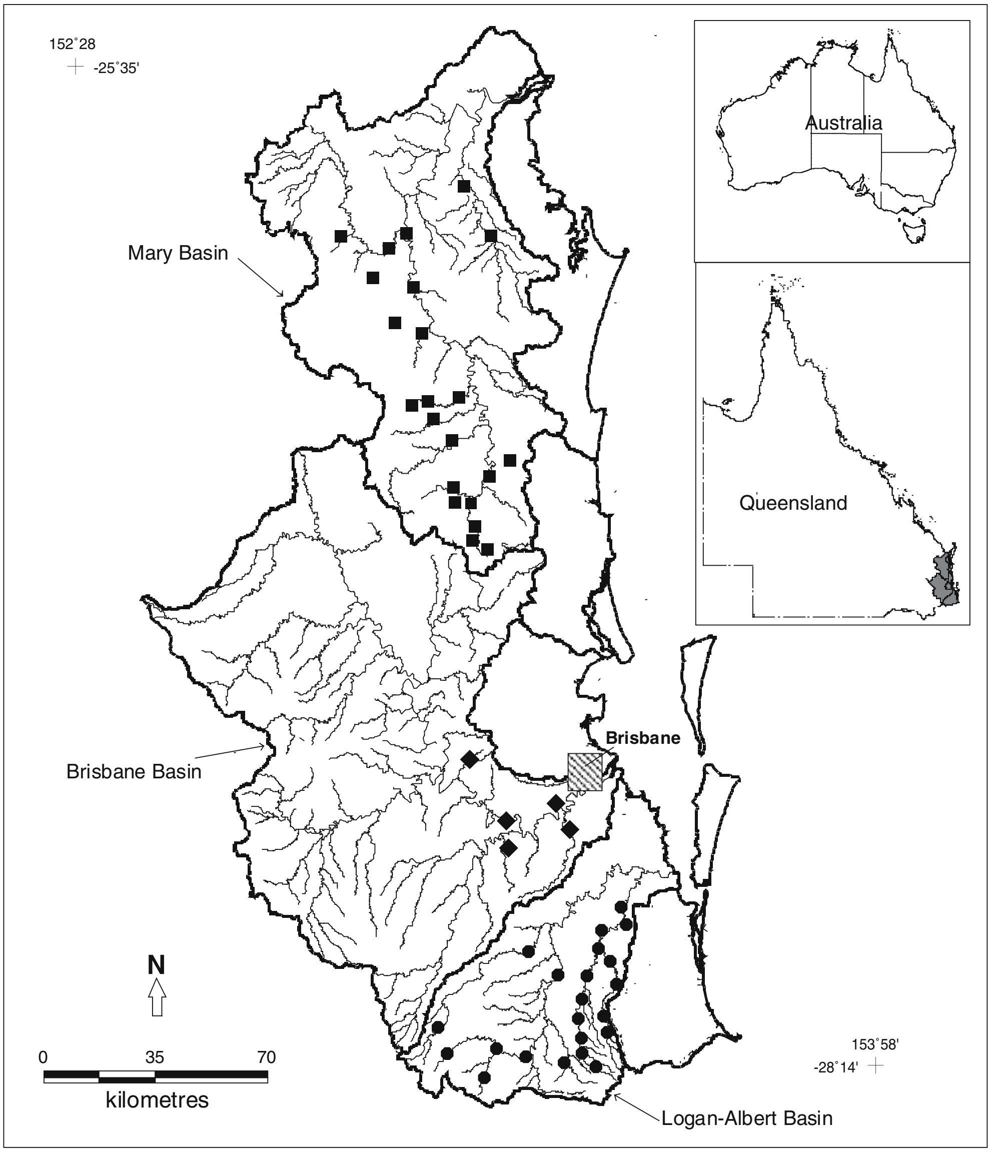

Figure 1. Location of the 53 study sites in three river basins in south-eastern Queensland. The inset shows the location of the study area in Queensland, Australia.

obtained from 53 sites sampled periodically (on 1–11 occasions) between 1994 and 1997.

Catchment-scale and local habitat variables

Catchment- and local-scale variables were selected based on a conceptual model of the relationship between processes at these two scales, and in addition were guided by the work of Davies et al. (2000) in the local region (see Table 1). In the selection process, catchment-scale variables that could potentially be modified by human influence were avoided. Local habitat variables focused on hydraulic, geomorphological and potentially ecologically important components of the habitat. A general approach was taken in the choice of local habitat descriptors because an objective of the study was to predict local habitat for different fish species, each potentially with differing habitat requirements.

Sampling of local habitat variables and fish

Sites were usually between 70 and 80 m of stream length and usually consisted of an entire meander wavelength or riffle-run-pool sequence. Catchment- and local-scale habitat variables were estimated for each site according to a standard protocol described in Pusey et al. (2004) (Table 1). Catchment descriptors for each site were estimated from 1:100,000 topographic maps using a digital planimeter or from Geographical Information Systems databases. Wetted stream width, mean water velocity and water depth were measured at a series of points located randomly throughout the site. Usually 40–60 random points were surveyed within each river reach. Substrate composition was estimated for one square metre around each survey point and allocated to each of seven substrate classes (Table 1) as a proportional representation. The abundance of submerged microhabitat struc-

Table 1. Catchment and local scale variables

Catchment scale variables	Local scale variables
Rainfall (rainMEAN)	Mean stream width (AVWIDTH)
Stream order (ORDER)	Mean stream depth (AVDEPTH)
Altitude (ELEVATION)	Maximum stream depth (MAXDEP)
Catchment area (CATAREA)	Mean stream flow rate (AVFLOW)
Latitude (LATDEC)	Maximum stream flow rate (MAXFLOW)
Longitude (LONDEC)	Mud substrate (MUD)[b]
Distance of site from source (DISTSOURCE)	Sand substrate (SAND)[b]
Distance of site from mouth[a]	Fine gravel substrate (FGRAVEL)[b]
Soil alkalinity/acidity[a]	Coarse gravel substrate (CGRAVEL)[b]
Soil infiltration rate[a]	Gravel substrate (FCG)[b]
Geology[a]	Cobble substrate (COBBLE)[b]
	Rock substrate (ROCKS)[b]
	Bedrock substrate (BEDROCK)[b]
	Macrophyte cover (MAC)[b]
	Leaf litter (LL)[b]
	Submerged overhanging vegetation (OHV)[b]
	Submerged marginal vegetation (SV)[b]
	Emergent vegetation (EV)[b]
	Large woody debris (LWD)[b]
	Small woody debris (SWD)[b]
	Filamentous algae (FA)[b]
	Undercut banks (UC)[c]
	Root masses (RM)[c]

[a] Not used in final models

[b] Mean percentage of site area

[c] Mean percentage of site wetted perimeter.

tures (aquatic macrophytes, filamentous algae, leaf litter, submerged vegetation (mainly grasses), emergent vegetation, submerged overhanging vegetation, large (>15 cm diameter) and small (1 cm < diameter < 15 cm) woody debris was also estimated at each survey point. In addition, the lineal extent (proportion of wetted perimeter) of undercut banks and root masses was estimated from multiple transect segments (every 10 m) along each bank. Average values (wetted width, depth and velocity), or average proportion of mean wetted site area (substrate composition and microhabitat structures) or stream bank (undercut banks and root masses) were then calculated for each site.

Fish assemblages at each site were intensively sampled using the procedures detailed in Pusey et al. (1998). Individual hydraulic units (i.e. riffles, runs or pools) within each site were sampled separately and data subsequently combined for the entire site. Each hydraulic unit was blocked upstream and downstream with weighted seine nets (11 mm stretched-mesh) to prevent fish movement into or out of the study area. The site was sampled using a combination of repeated pass electrofishing (Smith-Root model 12B Backpack Electroshocker) and supplementary seine netting until few or no further fish were collected (usually four electrofishing passes and two seine hauls were required to collect all fish present). The intensive sampling regime described here has been demonstrated to provide accurate estimates of species composition and abundances at wadable stream sites (Pusey et al., 1998) and intensive sampling of a single meander wavelength is the appropriate spatial scale at which to accurately and precisely characterise local fish assemblages in south-eastern Queensland streams and rivers (Kennard et al., 2006).

Modelling approach

Data analysis was carried out to determine which local habitat variables could be predicted from catchment-scale variables and whether the predictions would be accurate enough for use in fish presence–absence distribution models. The Assessment by Nearest Neighbour Analysis (ANNA) method (Linke et al., 2005) has the ability to make predictions on continuous data and has been used to predict the distribution of macroinvertebrate taxa among sites using catchment characteristics. This study employs ANNA as a method for predicting site habitat variables (instead of taxa) from catchment-scale variables. ANNA finds training sites that are most similar to a site of interest in terms of catchment characteristics (using a Euclidean dissimilarity index) and estimates each of the local variables of the given site as the mean of each of these variables for the selected training sites. ANNA was used instead of other methods like multiple regression because ANNA can predict a suite of variables at once and does not assume linearity or monotony in the models.

A subset of the data was created by randomly selecting one sample (a repeat visit) from each site. This subset, referred to as the training data, was used to develop the ANNA model. The site samples not included in the training data were referred to as the test data and used for model validation. ANNA employed the leave-one-out approach (Fielding & Bell, 1997) to make predictions on the withheld samples. Thus, for a given test site, the values of local habitat variables from other sites were used to calculate this sites' local habitat values. Even though the test data were not spatially independent from the training data, the use of the leave-one-out approach for making predictions on the test data would have avoided optimistic assessment of the accuracy of the ANNA predictions.

The analysis involved predicting local habitat variables on test samples and then using the variables that could be predicted as candidate variables in the development of fish presence–absence models based on training data. For each fish species, distribution predictions were made on test samples based on actual local habitat variables and on ANNA predicted habitat local variables. Each of the predictions was assessed to determine how well the model could discriminate between occupied and unoccupied sites. The predictive performances of the models on test samples using observed and predicted local variables were compared to determine if there was a difference in model performance between the two sets of predictor variables.

Assessment of the predictive performance of presence–absence models

Although prediction error is a more intuitive measure of model performance, it is not appropriate for presence–absence data because it is influenced by the prevalence of the presence or absence records (Fielding & Bell, 1997; Franklin, 1998). For instance, a model that predicts a species that occupies 5% of the sample sites as being absent everywhere will have a low prediction error simply because the species is not well represented in the data set. A model performance measure should incorporate the ability of the model to correctly predict positive cases (sensitivity), i.e. species presence, and the ability of the model to correctly predict negative cases (specificity), i.e. species absence (Fielding and Bell, 1997). The ability of a model to discriminate between occupied and unoccupied sites means that the predictions from the model are a good index of species distribution even if the actual predicted values do not represent true probability of occurrence (Pearce and Ferrier, 2000).

The receiver operating characteristic (ROC) approach (Hanley & McNeil, 1982) is an increasingly popular way of evaluating the performance of diagnostic and predictive test systems (Hanley & McNeil, 1983; Centor & Schartz, 1985; DeLong et al., 1988) such as species presence–absence distribution models (Fielding & Bell, 1997; Pearce & Ferrier, 2000). A ROC plot is obtained by plotting model sensitivity vs. (1 – specificity) (Swets & Pickett, 1982). The area under the curve (AUC) of a ROC plot corresponds to the probability of correctly predicting positive and negative cases (Green & Swets, 1966). The AUC of a ROC plot ranges between 0.5 (random) and 1 (perfect model) (Fielding & Bell, 1997). A value of 0.8 for the AUC means that for 80% of the time a random selection from the positive group will have a score greater than a random selection from the negative class (Fielding & Bell, 1997).

For a given model, the observed and predicted values of species occurrence can be used by the ROC program of Atkinson and Mahoney (2001) to calculate AUC of a ROC plot and to perform a one sided z-test to determine whether the AUC is significantly greater than 0.5. The program can also compare the performance of different models based on the same observations. Significance testing for the comparison of the predictive performance of different models is corrected for the correlated nature of data using the DeLong et al. (1988) method (Atkinson & Mahoney, 2001). The program was used to calculate model predictive performance and for the comparisons of model performance in the stepwise model-building process for fish distribution models.

Local habitat prediction

An ANNA model was developed to predict local habitat variables using the training data set. The development of the ANNA model described here follows the methods in Linke et al. (2005) with the modification that the value of local habitat variables were predicted for a given site instead of taxa occurrence. For each local habitat variable the predictive success was measured as the R^2 from the regression of the predicted and observed values. Geological and geomorphological variables were available for use as predictors in the ANNA model but their inclusion in earlier trials did not improve the predictive performance of the model. Therefore, only the simple predictor variables (mean rainfall, catchment area, stream order, latitude, longitude, distance from source and altitude) were used in model construction. In the ANNA model, the prediction residuals were calculated as the difference between the range-standardized predicted and observed values of a local habitat variable. The local habitat variable was range standardized as shown below.

Standardized value = (input value−minimum of input variable)/range of input variable.

The predicted value was standardized using the minimum and range values used to standardize the observed values of the respective local habitat variables. A statistical threshold of R^2 greater than 0.2 for predictions was used to select variables for which the ANNA model had reasonable predictive performance.

Fish distribution modelling

The local habitat variables that were predicted by ANNA with an R^2 greater than 0.2 in the test data set were used as candidate predictor variables in logistic regression models for fish species distribution. An attempt was made to develop distribution models for each of the nine fish species in this study (see results), all of which are relatively common and widespread in south-eastern Queensland rivers and streams (Pusey et al., 2004).

Since the response was a fish species presence–absence variable the logistic regression models assumed binomial error. Logistic regression models for species distribution were developed using the forward selection procedure (Draper & Smith, 1966) with a significance level of 0.05 for variable entry. The training data set was small (53) and the models had to be conservative; therefore, efforts were made to avoid fitting the models with too many variables so that models become training data specific and not useful for independent data. If a model is overfitted its parameter estimates or standard errors may become extremely large and predictive performance on independent data is reduced (Hosmer & Lemeshow, 1989). Therefore, a lack of improvement in model predictive performance or the appearance of large parameter estimates or standard errors upon the addition of a new variable to the model were also used as stopping criteria for model building in order to avoid overfitting the models. Making predictions on the same data that the model is derived from leads to optimistic measures of model performance (Fielding & Bell, 1997). Thus, a model specified from each stage in the forward selection procedure was fitted to 200 bootstrap samples containing 53 observations (sampled with replacement) (James & McCulloch, 1990; Efron & Tibshirani, 1993) and used to predict the training dataset in order to obtain independent prediction estimates. A seed was specified for the bootstrapping procedure so that analysis could be repeated using the same 200 bootstrap samples. The mean of the predictions for each training sample over the 200 bootstrap samples was taken as the probability of occurrence of a given fish species. The bootstrapped predictions were used to test whether the addition of a variable improved model predictive performance at each stage of the forward selection procedure. Model performance improved where the AUC of a ROC plot for a proposed model was significantly (significance level of 0.05) greater than that of the previous model (i.e. submodel). Where model performance did not improve the previous model was selected as the distribution model for the species. For each species the selected model was fitted to the 200 bootstrap samples and used to make bootstrapped predictions with the training data, test data (withheld samples) based on observed values of the selected local variable, and test data based on the predicted values of the selected local variable.

Comparison of the predictive performance of fish distribution models on test data based on actual and predicted local habitat variables

The mean of the bootstrap predictions on each of the withheld samples was taken as the probability of occurrence of a given fish species. ANOVA was used to test whether model predictive performance differed with prediction type (fish distribution predictions on withheld replicates based on observed habitat local variables and fish distribution predictions based on ANNA predicted habitat local variables).

Results

The ANNA method was provided with latitude (LATDEC), catchment area (CATAREA), mean annual rainfall (rainMEAN), stream order (ORDER), longitude (LONDEC), distance from the source (DISTSOURCE) and elevation (ELEVATION) as candidate predictor variables. The first three variables were not used in the actual ANNA local habitat variable prediction model because they were not selected as predictor variables by the ANNA model as shown by their zero weights in Table 2 (see Linke et al. (2005) for details on

Table 2. The ANNA weights given to each predictor variable along a given dimension in the prediction of local habitat variables; see Linke et al. (2005) for details on assigning weights for large-scale variables

Dimensions	rainMEAN	ORDER	LATDEC	LONDEC	CATAREA	DISTSOURCE	ELEVATION
1	0	1.24898	0	0	0	−3.66410	−3.31292
2	0	1.17926	0	0	0	1.70245	0
3	0	−2.33892	0	−1.85608	0	0	−1.74748

Variables with a weight of zero along the three dimensions are not used in the prediction of local variables.

Table 3. Local habitat variables that were successfully predicted (i.e. $R^2 > 0.2$) on the test samples by the ANNA model using catchment characteristics as predictor variables

Local variable	R^2 on test samples
AVWIDTH	0.37
SAND	0.22
COBBLE	0.28
ROCKS	0.65
LWD	0.42

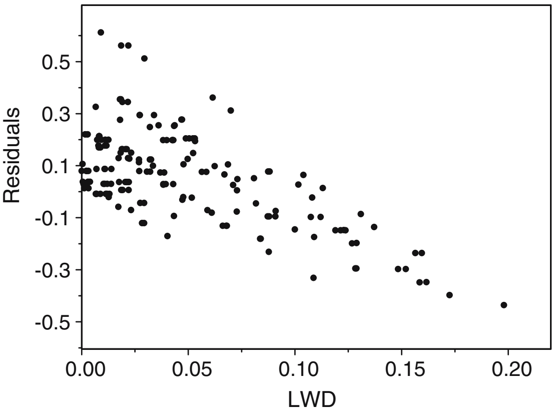

Figure 2. ANNA prediction residuals for the LWD showing a tendency for high positive residuals (over-estimates) at the low end of the gradient and high negative residuals (under-estimates) at the high end of the gradient. The residuals were calculated as the difference between the range-standardized (as shown in the methods) predicted and observed values of the LWD variable.

assigning variable weights). The ANNA model was able to predict ROCKS, COBBLE, SAND, LWD and AVWIDTH with an R^2 greater than 0.2 (Table 3). The plot of the ANNA prediction residuals for the LWD variable (Fig. 2) illustrates the trend of positive residuals (over-estimates) at the low end of the gradient and negative residuals (under-estimates) at the high end of the gradient that was evident in the residual plots for ROCKS, SAND, COBBLE and AVWIDTH. The local habitat variables that were predicted by ANNA with an R^2 greater than 0.2 were taken as candidate predictor variables for modelling distributions of the individual fish species using logistic regression based on training data.

Logistic regression models were developed for the individual distributions of the fish species *Retropinna semoni*, *Melanotaenia duboulayi*, *Tandanus tandanus*, *Mogurnda adspersa*, *Pseudomugil signifer*, *Gobiomorphus australis* and *Philypnodon grandiceps*. No models could be selected for *Hypseleotris galii* and *Ambassis agasizii* using the model building criteria employed in this study. For each fish species, only a single local habitat variable could be included in the logistic regression models without over-fitting. The predictive performances (as measured by the area under the curve (AUC) of a ROC plot) of each species distribution model on the training samples, test samples with observed values of local variables and test samples with predicted local variables are shown in Table 4. According to the one sided

Table 4. The predictive performance of logistic regression models for fish species distribution as measured by the area under the curve of a ROC plot (AUC) for each species and for each prediction type (training data predictions, test data predictions based on the values of observed local habitat variables and test data predictions based on predicted values of local habitat variables)

Fish species	Local habitat variables selected in fish distribution model	AUC for bootstrapped predictions on Training data using actual local variables	AUC for bootstrapped predictions on Test data using actual local variables	AUC for bootstrapped predictions on Test data using predicted local variables
Retropinna semoni	COBBLE	0.82	0.67	0.57
Mogurnda adspersa	ROCKS	0.64	0.68	0.56
Melanotaenia duboulayi	AVWIDTH	0.84	0.64	0.67
Tandanus tandanus	AVWIDTH	0.90	0.76	0.75
Pseudomugil signifer	COBBLE	0.81	0.59	0.80
Gobiomorphus australis	LWD	0.74	0.84	0.95
Philypnodon grandiceps	LWD	0.80	0.74	0.68

Also shown are the local habitat variables selected by logistic regression models for each fish species.

z-test (performed for each AUC calculation) all the predictive performances in Table 4 were significantly greater than an AUC of 0.5. The AUC value for *Gobiomorphus australis* for test data using predicted local habitat values was 0.95, meaning that 95% of the time, a random selection from the species present sites had a higher predicted probability of occurrence than a random selection from the species absent sites. For some species such as *Gobiomorphus australis* the logistic regression model predictions of fish distribution in withheld samples made using predicted habitat variables were superior to the predictions derived from observed values of the local habitat variable (Table 4). A two factor ANOVA showed that neither prediction type (predictions based on observed values in withheld samples and predictions based on ANNA predicted habitat local variables in the withheld samples) (p-value = 0.85) nor fish species (p-value = 0.13) affected model performance.

Discussion

This study showed that, with the use of a few catchment-scale variables (stream order (ORDER), longitude (LONDEC), distance from the source (DISTSOURCE) and elevation (ELEVATION)) that are easily derived from maps, the ANNA method was able to predict five local habitat variables in the test data set with an R^2 greater than 0.2 (Table 3). Similarly, Jeffers (1998) was able to predict the occurrence of site features based on only four map-derived variables (altitude, slope, distance from source and height of source above sea level). However, unlike Jeffers (1998) where the probability of occurrence of a site feature was predicted, we successfully predicted the value of local variables on a continuous scale from a few catchment characteristics that are relatively cheap to acquire. Given that ANNA estimates local habitat values from the nearest neighbours, the pattern of positive residuals (over-estimates) at the low end of a local habitat gradient and negative residuals (under-estimates) at the high end of the gradient (as in the LWD example, Fig. 2) should be reduced by using a larger training dataset so that more neighbours that are similar are available. The successful prediction of local habitat variables from catchment-scale variables concurs with studies that have suggested that the local habitat is influenced by factors operating at larger spatial scales (Frissell et al., 1986; Imhof et al., 1996; Richards et al., 1996; Davies et al., 2000).

In future studies, the accuracy of the ANNA habitat predictive model may be improved by handling the temporal variation in the local habitat variables in other ways. For instance, inclusion of winter and spring samples in the training and testing data sets is likely to have introduced temporal variability, which cannot be accounted for by the predictor variables used the ANNA model (the predictors were all "static" in time). In addition to exaggerating the magnitude of residuals for the local habitat predictions, the temporal effects might have also obscured the relationships among some local habitat variables and large-scale variables. Future modelling will account for the effects of temporal variation in the values of local habitat variables by either including appropriate antecedent predictor variables, or by constructing separate seasonal models.

The influence of human activities on local habitat characteristics was not determined in this study because only sites close to natural condition were used. The ability to predict local habitat variables in the absence of human impacts is useful because the predicted values of variables can be compared with observed values so that the degree of habitat degradation at a site can be assessed. However, the prediction accuracy of local habitat variables would need to be improved (e.g. accounting for seasonal variation in habitat variables) before using them to assess the habitat condition of sites. As part of further research, improved local habitat predictions will be used to assess habitat condition in impacted sites.

Logistic regression models fitted with a single local-scale habitat variable produced accurate predictions of fish species distributions (Table 4). These habitat attributes may be of direct importance for some species to satisfy critical life history requirements or may represent local surrogates or correlates for other large-scale physical factors that influence fish distributions. Submerged physical structures such as rocks, cobbles and woody debris are commonly used as a source of refuge, spawning substrate and are likely to support invertebrate food resources for many fish species present in south-east Queensland, including those species considered here (Merrick & Schmida, 1984; Pusey et al., 2004). The distribution and abundance of these local habitat attributes is in turn likely to be influenced by physical processes operating at larger spatial scales (Frissell et al., 1986; Imhof et al., 1996; this study). Large-scale catchment features such as elevation, proximity to river mouth have also been shown to influence fish distributions within river systems (Pusey & Kennard, 1996; Belliard et al., 1997; Pusey et al., 2004) and can been used to accurately predict local fish species composition in south-east Queensland rivers (Kennard et al., 2006) and elsewhere (Joy & Death, 2002; Olden & Jackson, 2002). Although fish species respond to large-scale variables, Bond & Lake (2003) showed that fish distributions were more strongly associated with the presence of habitat structure at the scale of metres suggesting that, in disturbed rivers, fish abundances are limited by the low availability of habitat at these small spatial scales. Thus, sensitivity of fish species to local habitat variables may explain the high predictive performance of species distribution models found in this study.

The usefulness of predicted values of local habitat variables for the prediction of fish presence–absence distribution was assessed on withheld samples by comparing the performance of fish distribution models where the predictions were based on observed values of local habitat variables and where predictions were based on ANNA predicted values of local habitat variables (Table 4). Neither prediction type, nor fish species had an effect on the predictive performance of logistic regression species distribution models (two factor ANOVA). Therefore, local habitat variables that were predicted by the ANNA model with R^2 greater than 0.2 were accurate enough to produce fish presence–absence predictions comparable to fish distribution predictions based on observed local habitat variables.

The predictive performance of fish models based on predicted local habitat variables was dependent on the interaction among errors in the predicted local habitat variables and errors in the distribution of the species along the local habitat variable. In some cases the estimates of the local habitat variables based on catchment characteristics resulted in fish distribution predictions superior to distribution predictions based on observed values of habitat variables. For instance, there was an improvement in *Gobiomorphus australis* prediction when predicted values of LWD were used instead of observed values of LWD (Table 4). *Gobiomorphus australis* had low occurrence in the low part of the LWD gradient but occurred more frequently at higher values of LWD (Fig. 3). In this case, under-estimates of LWD at the high end of the LWD gradient did not adversely affect the predictive performance of the fish distribution model because the predicted LWD values were still in the range of preferred habitat values for the fish. In the midrange of the LWD gradient the prediction, residuals were lower than at the extreme ends (Fig. 2) and thus maintained good predictive performance of *Gobiomorphus australis* distribution. The overall improvement of *Gobiomorphus australis* predictions suggest that the predicted LWD values might have captured the long-term values of LWD (as determined by the large-scale variables) to which the fish species might be responding, instead of the instantaneous observed values of the local habitat variables. *Retropinna semoni* exemplified cases where the predictive performance of a fish distribution model based on ANNA predicted values for a local habitat variable was inferior to the performance based on observed values of a local habitat variable. The observed occurrence of *Retropinna semoni* increased sharply from the low end of the COBBLE gradient then plateaued from

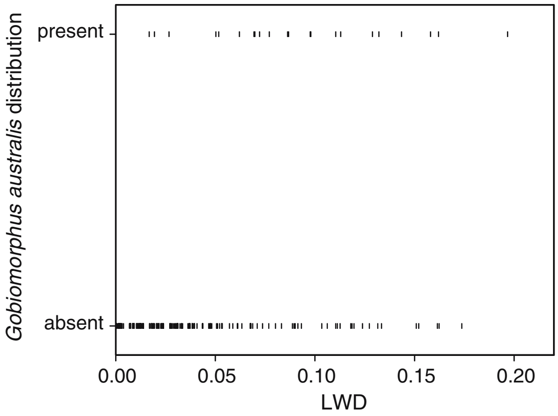

Figure 3. The distribution of *Gobiomorphus australis* along a gradient of the observed values of LWD showing that the fish species is not present at very low values of LWD but becomes relatively more common at higher values of LWD.

the mid-to-high range of the COBBLE gradient. The ANNA model overestimated the values of the COBBLE variable for samples in the low end of the COBBLE gradient giving similar estimates of COBBLE value to samples in the low and mid-range of the COBBLE variable. Consequently, the logistic regression model for *Retropinna semoni* predicted similarly high probability of occurrence in the low and mid-to-high range COBBLE samples leading to the poor distribution predictions of *Retropinna semoni* based on ANNA estimates of undisturbed values of COBBLE. However, given that some fish distribution models had good predictive performance based on ANNA predicted values of local habitat variables, a determination can be made about whether those fish species could actually live at a site if the undisturbed habitat was available.

There are many conservation implications in being able to predict local habitat and fish distribution in the absence of disturbance (Maddock, 1999; Davies et al., 2000; Olden & Jackson, 2002). Information on the degree of habitat degradation at a site could assist in the accurate diagnosis of the mechanisms responsible for observed deviations in biotic composition from that expected from species distribution models. Combining information derived from predictive models of local habitat structure and biotic composition should also enable managers to accurately identify those sites that may be in need of habitat restoration or remediation.

References

Armitage, P. D., R. J. M. Gunn, F. M. T., J. F. Wright & D. Moss, 1987. The use of prediction to assess macroinvertebrate response to river regulation. Hydrobiologia 144: 25–32.

Atkinson, B. & D. Mahoney, 2001. S-Plus ROC functions. In Mayo Foundation for Medical Education and Research.

Belliard, J., P. Boet & E. Tales, 1997. Regional and longitudinal patterns of fish community structure in the Seine River basin, France. Environmental Biology of Fishes 50: 133–147.

Bond, N. R. & P. S. Lake, 2003. Characterizing fish-habitat associations in streams as the first step in ecological restoration. Austral Ecology 28: 611–621.

Centor, R. M. & J. S. Schartz, 1985. An evaluation of methods for estimating the area under the Receiver Operating Characteristic (ROC) curve. Medical Decision Making 5: 149–156.

Davies, N. M., R. H. Norris & M. C. Thoms, 2000. Prediction and assessment of local stream habitat features using large-scale catchment characteristics. Freshwater Biology 45: 343–369.

De Boer, D. H., 1992. Hierarchies and spatial scale in process geomorphology: A review. Geomorphology 4: 303–318.

DeLong, E. R., D. M. DeLong & D. L. Clarke-Pearson, 1988. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. Biometrics 44: 837–845.

Draper, N. R. & H. Smith, 1966. Applied Regression Analysis. John Wiley and Sons, Inc, New York.

Efron, B. & R. J. Tibshirani, 1993. An Introduction to the Bootstrap. Chapman and Hall, New York.

Fielding, A. H. & J. F. Bell, 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. Environmental Conservation 24: 38–49.

Franklin, J., 1998. Predicting the distribution of shrub in southern California from climate and terrain-derived variables. Journal of Vegetation Science 9: 733–748.

Frissell, C. A., W. J. Liss, C. E. Warren & M. D. Hurley, 1986. A hierarchical framework for stream habitat classification: Viewing streams in a watershed context. Environmental Manangement 10(2): 199–214.

Gafny, S., M. Goren & A. Gasith, 2000. Habitat condition and fish assemblage structure in a coastal mediterranean stream (Yarqon, Israel) receiving domestic effluent. Hydrobiologia 422: 319–330.

Gorman, O. T. & J. R. Karr, 1978. Habitat structure and stream fish communities. Ecology 59: 507–515.

Green, D. & J. A. Swets, 1966. Signal Detection Theory and Psychophysics. John Wiley and Sons, New York.

Hall, L. W., R. P. Morgan, E. S. Perry & A. Waltz, 2002. Development of a provisional physical habitat index for Maryland freshwater streams. Environmental Monitoring and Assessment 77: 265–291.

Hanley, J. A. & B. J. McNeil, 1982. The meaning and use of the area under receiver operating characteristic curve. Radiology 143: 29–36.

Hanley, J. A. & B. J. McNeil, 1983. A method for comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 148: 839–843.

Harper, D. & M. Everard, 1998. Why should the habitat-level approach underpin holistic river survey and management? Aquatic Conservation-Marine and Freshwater Ecosystems 8: 395–413.

Hosmer, D. W. & S. Lemeshow, 1989. Applied Logistic Regression. John Wiley & Sons, Brisbane.

Imhof, J. G., J. Fitzgibbon & W. K. Annable, 1996. A hierarchical evaluation system for characterizing watershed ecosystems for fish habitat. Canadian Journal of Fisheries and Aquatic Sciences 53: 312–326.

James, F. C. & C. E. McCulloch, 1990. Multivariate analysis in ecology and systematics: Panaceae or pandora's box. Annual Review of Ecology and Systematics 21: 129–166.

Jeffers, J. N. R., 1998. Characterization of river habitats and prediction of habitat features using ordination techniques. Aquatic Conservation-Marine and Freshwater Ecosystems 8: 529–540.

Joy, M. K. & R. G. Death, 2002. Predictive modelling of freshwater fish as a biomonitoring tool in New Zealand. Freshwater Biology 47: 2261–2275.

Kauffman, J. B., R. L. Beschta, N. Otting & D. Lytjen, 1997. An ecological perspective of riparian and stream restoration in the western United States. Fisheries 22: 12–24.

Kennard, M. J., B. J. Pusey, A. H. Arthington & S. J. Mackay, 2006. Development and application of a predictive model of freshwater fish assemblage composition to evaluate river health in eastern Australia. Hydrobiologia. This Volume.

Knighton, D., 1984. Fluvial Forms and Processes. Edward Arnold, London, UK.

Linke, S., R. Norris, D. P. Faith & D. Stockwell, 2005. ANNA: A new prediction method for bioassessment programs. Freshwater Biology 50: 147–158.

Maddock, I., 1999. The importance of physical habitat assessment for evaluating river health. Freshwater Biology 41: 373–391.

McCullagh, P. & J. A. Nelder, 1989. Generalized Linear Models. Chapman and Hall, London, New York.

Merrick, J. R. & G. E. Schmida, 1984. Australian Freshwater Fishes: Biology and Management. Griffin Press Ltd, Netley, South Australia.

Olden, J. D. & D. A. Jackson, 2002. A comparison of statistical approaches for modelling fish species distributions. Freshwater Biology 47: 1976–1995.

Pearce, J. & S. Ferrier, 2000. An evaluation of alternative algorithms for fitting species distribution models using logistic regression models. Ecological Modelling 128: 127–147.

Poff, N. L., 1997. Landscape filters and species traits: Towards mechanistic understanding and prediction in stream ecology. Journal of the North American Benthological Society 16: 391–409.

Poff, N. L. & J. D. Allan, 1995. Functional-Organization of Stream Fish Assemblages in Relation to Hydrological Variability. Ecology 76: 606–627.

Poff, N. L. & J. V. Ward, 1990. The physical habitat template of lotic systems: Recovery in the context of historical pattern of spatio-temporal heterogeneity. Environmental Management 14: 629–645.

Pusey, B. J. & M. J. Kennard, 1996. Species richness and geographical variation in assemblage structure of the freshwater fish fauna of the Wet Tropics region of northern Queensland. Marine and Freshwater Research 47: 563–573.

Pusey, B. J., M. J. Kennard & A. H. Arthington, 2004. Freshwater Fishes of North-eastern Australia. CSIRO Publishing, Collingwood.

Pusey, B. J., M. J. Kennard, J. M. Arthur & A. H. Arthington, 1998. Quantitative sampling of stream fish assemblages: Single- versus multiple pass electrofishing. Australian Journal of Ecology 23: 365–374.

Richards, C., R. J. Haro, B. L. Johnson & G. E. Host, 1997. Catchment and reach-scale properties as indicators of macroinvertebrate species traits. Freshwater Biology 37: 219–230.

Richards, C., L. B. Johnson & G. E. Host, 1996. Landscape-scale influences on stream habitats and biota. Canadian Journal of Fisheries and Aquatic Sciences 53(Suppl. 1): 295–311.

Schlosser, I. J., 1995. Critical landscape attributes that influence fish population dynamics in headwater streams. Hydrobiologia 303: 71–81.

Schlosser, I. J. & P. L. Angermeier, 1995. Spatial variation in demographic processes of lotic fishes: Conceptual models, empirical evidence, and implications for conservation. American Fisheries Society Symposium 17: 392–401.

Schumm, S. A. & R. W. Lichty, 1965. Time, space and causality in geomorphology. American Journal of Science 263: 110–119.

Swets, J. A. & R. M. Pickett, 1982. Evaluation of Diagnostic Systems. Academic Press, New York.