The Scottish Government
Riaghaltas na h-Alba

Scottish Marine and Freshwater Science Report

Volume 6 Number 6

Development of a Model for Predicting Large Scale Spatio-Temporal Variability in Juvenile Fish Abundance from Electrofishing Data

C P Millar, K J Millidine, S J Middlemas & I A Malcolm

marinescotland

Scottish Marine and Freshwater Science Vol 6 No 6

Development of a Model for Predicting Large Scale Spatio-Temporal Variability in Juvenile Fish Abundance from Electrofishing Data

C P Millar, K J Millidine, S J Middlemas & I A Malcolm

Published by Marine Scotland Science

ISSN: 2043-7722
DOI: 10.7489/1616-1

Marine Scotland is the directorate of the Scottish Government responsible for the integrated management of Scotland's seas. Marine Scotland Science (formerly Fisheries Research Services) provides expert scientific and technical advice on marine and fisheries issues. Scottish Marine and Freshwater Science is a series of reports that publishes results of research and monitoring carried out by Marine Scotland Science. It also publishes the results of marine and freshwater scientific work that has been carried out for Marine Scotland under external commission. These reports are not subject to formal external peer-review.

This report presents the results of marine and freshwater scientific work carried out by Marine Scotland Science.

Marine Scotland Science
Freshwater Laboratory
Faskally
Pitlochry
PH16 5LB

Copies of this report are available from the Marine Scotland website at http://www.gov.scot/marinescotland

Development of a Model for Predicting Large Scale Spatio-Temporal Variability in Juvenile Fish Abundance from Electrofishing Data

Colin P Millar, Karen J Millidine, Stuart J Middlemas & Iain A Malcolm

Marine Scotland Science, Freshwater Laboratory
Faskally, Pitlochry, PH16 5LB

Executive Summary

Models of juvenile salmonid abundance are required to inform electrofishing based assessment approaches and potentially as an intermediate step in scaling conservation limits from data rich to data poor catchments. This report describes an approach for modelling large-scale spatio-temporal variability in fish densities using GIS derived covariates. The technical challenges, modelling approaches and software developed during the project are described. The utility of the approach was illustrated by fitting models to data on Atlantic salmon fry.

Multi-pass electrofishing data was collated from fisheries trusts, fisheries boards, SEPA and Marine Scotland. Covariates describing spatial, temporal and habitat variability were obtained for each sampling event. A two stage likelihood based modelling approach was developed. Firstly, capture probability was modelled in relation to covariates. Secondly, densities were modelled in relation to covariates, conditional on estimated capture probabilities. A software package was developed for the R statistical programming environment to perform the analysis.

The modelling approach was illustrated using a case study of Atlantic salmon fry. Capture probability varied strongly with data provider, time of sampling (Day of Year: DoY), location and year and to a lesser extent with distance to sea, channel width and gradient. The risks of assuming constant capture probability were demonstrated by comparing modelled density estimates with those obtained by assuming a constant mean capture probability. Assumptions of constant capture probability resulted in considerable spatial bias in density estimates, but could also generate temporal bias. This illustrates the continued requirement for multi-pass electrofishing data and the

constraints on the use of single-pass or timed data for quantitative assessments in the absence of adequate calibration.

Fry densities varied with catchment, and with habitat variables: distance to sea, altitude, upstream catchment area and to a lesser extent channel width, gradient, DoY and landuse. The potential of spatial covariates (catchment and river connectivity) to indicate deviation from an average national model and thus describe the relative performance of catchments or sites is illustrated as a potential assessment approach for management.

It is recommended that future work focusses on (1) quality control of site location data and identification of non-representative electrofishing samples (e.g. fishing only marginal areas in wide rivers) through interaction with data providers and (2) further development of software to allow simultaneous consideration of multiple species (salmon and trout) and life stages (fry and parr) together with river network effects across multiple catchments.

Introduction

Large scale models of juvenile fish abundance are required to understand and predict spatial variability in fish productivity. In fisheries management, such models can inform the development of juvenile assessment tools from which to interpret electrofishing data (Godfrey, 2005; SNIFFER, 2011) or provide an intermediate step in scaling stock-recruitment relationships between data rich and poor catchments for the development of conservation limits (Wyatt and Barnard, 1997).

Fish productivity is known to vary at a range of spatial scales depending on water quality, food availability, hydraulic and sedimentary characteristics (Fausch *et al.*, 1988; Armstrong *et al.*, 2003). However, not all these variables can be adequately characterised at large spatial scales suitable for the development of regional or national models. Under such circumstances, it is necessary to substitute process based predictors for mechanistically plausible Geographical Information System (GIS) derived surrogates to make spatial predictions of abundance. These surrogates typically take the form of landscape characteristics (e.g., channel slope) that can be defined using a GIS, thereby allowing covariates to be readily obtained for sampling sites and predictions to be made for un-monitored locations.

In addition to the issues of scale and environmental characterisation, the development of juvenile density models is affected by technical challenges in two main areas: Firstly, the need to obtain density estimates from electrofishing observations, where catch probability likely varies with a range of factors including habitat, sampling equipment, procedures and personnel; and secondly, the need to model spatial variability in density where this is influenced by habitat, but is also spatially correlated.

Recent attempts to model spatial variability in juvenile fish abundance at large spatial scales have attempted to combine models of capture probability and spatial prediction using hierarchical Bayesian models (Wyatt, 2002, 2003; Rivot *et al.*, 2008). Hierarchical Bayesian models can readily accommodate complicated model structures and are well suited for modelling electrofishing data. As such, these models provide joint distributions of the estimates of capture probability and abundance, integrating uncertainty across multiple sites. However, such complex models often require considerable time (days) to fit. In addition, because they are tailored for specific analyses, model exploration and comparison is time consuming and thus limiting. Likelihood based alternatives could offer significant benefits in terms of rapid fitting. Furthermore where generalised additive models can be used, this provides model flexibility (Wyatt *et al.,* 2007b) and ease of implementation using standard software e.g., mgcv (Wood, 2011).

Unfortunately, between site variability in densities and capture probabilities cannot be estimated simultaneously using GAMs because the resulting likelihood is not an exponential family distribution (Wood, 2006). However, where capture probability is estimated separately, GAMs can be used to model density. Consequently, many previous approaches have estimated capture probability independently for individual site visits (Otis *et al.*, 1978; Lanka *et al.*, 1987; Huggins and Yip, 1997) or at the other extreme assumed a constant capture probability (Bohlin *et al.*, 2001). Approaches that focus on individual site visits can result in poorly defined capture probabilities (and thus density estimates) or provide very uncertain estimates of density in the case where no fish are caught. Conversely, an assumption of constant capture probability across all sites is potentially an over simplification and could lead to an under-estimate of uncertainty or to model bias where estimates of density are fed into subsequent spatial habitat models. Improvements in density estimates could be made if capture probability were to be modelled in a flexible framework where it was allowed to vary with mechanistic covariates.

In contrast to terrestrial and marine spatial models, data collected on rivers have specific spatial structuring that is not directly related to the distance between sampling points, but instead to the combined effects of geographical (as the crow flies or Euclidian distance) and network distance. There have been a large number of statistical developments in recent years that could improve models of fish densities on river networks. For example, geostatistical approaches (Cressie *et al.,* 2006; Peterson *et al.*, 2013), or the use of Gaussian Markov random fields (Wyatt, 2003; Rue and Held, 2005). However, at present, there is no single software package that allows ready specification of models that could include variation at different spatial scales, on river networks, over time and in relation to landscape covariates. Such models would be desirable in the context of fish abundance studies.

This report describes the development of a two stage modelling approach that can be used to characterise, understand and predict spatio-temporal variability in fish abundance at the Scottish scale. This was achieved by developing models to (1) predict capture probability from removal electrofishing data and (2) to understand and predict spatio-temporal variability in fish abundance using GIS derived landscape characteristics and fish abundances estimated from the catch probability model.

As a first step, development and application of the models was undertaken using a single species and life stage: Atlantic salmon fry.

The report is structured according to the following specific objectives:

1. Identify, request and collate electrofishing data from internal (MSS) and external sources including SFCC and SEPA.
2. Identify a set of GIS covariates that could plausibly influence fish abundance or capture probability and calculate these for each electrofishing site. Describe the processes involved and technical difficulties associated with the generation of covariates.
3. Develop models for estimating capture probability from depletion electrofishing data.
4. Develop models for estimating spatio-temporal variability in salmonid abundance.
5. Produce an R software package to fit the models identified in 4 & 5 and describe the process of model specification and selection.

6. Demonstrate model application using a case study of Atlantic salmon fry and interpret findings.
7. Make recommendations for future data collection, data processing, software development and modelling.

Data Availability

Given variability in catch efficiency (Borgstrom and Skaala, 1993; Niemela *et al.*, 2000), only multi-pass data were collated for this report. Data was obtained from the Scottish Fisheries Coordination Centre (SFCC) database, Marine Scotland Science (MSS) FishObs database and in spreadsheet format from Scottish Environment Protection Agency (SEPA) and Caithness District Salmon Fishery Board (Alan Youngson). Additional MSS data were compiled in spreadsheet formats where not already entered on the FishObs database. Sites that were known to be stocked (as identified by the SFCC stocking code or knowledge of MSS staff) were excluded to simplify modelling requirements. Because many data sources do not reliably obtain ages from scale reading, electrofishing data was resolved to life-stage (fry or parr) rather than individual age class. Throughout this report 'Site' is used to define a location visited for electrofishing; 'Site Visit' identifies individual visits to a Site. The starting dataset contained 2511 sites and 5468 site visits. Only those data collected between 1997 and 2013 and between the 1st June and 21st November were used in subsequent analyses (2367 sites, 4748 site visits) because few of the collated data lay outside of these time periods and thus there was insufficient information to inform large scale spatio-temporal models. The compiled data was also visually assessed in relation to covariates and outliers removed leaving 2353 sites and 4648 site visits. Because preliminary model fits suggested implausible relationships between fish density and high values of the covariates "Channel Width" and "Upstream Catchment Area", a manual check of raw data was performed on site visits with a width of >25m and an upstream catchment area of >500km^2. This check suggested a number of circumstances where only partial electrofishing surveys (e.g. peripheral areas or single channels in braided rivers) were carried out on larger rivers. Given the potential bias of these data (28 Site visits to 27 sites) they were also excluded from the analysis. Finally, any sites located above impassable barriers were removed from the dataset, so that the final dataset only contained locations that were accessible to salmon, reducing the final dataset to 3743 site visits from 1875 sites.

The spatial and temporal coverage of multi-pass electrofishing data in the final dataset is shown in Figure 1. In general, spatial coverage increased between 1997 and 2013, the exception being North West Scotland, where data coverage was better in earlier years. There was a general paucity of data from the islands, with no data available from Orkney or Shetland. A schematic showing the temporal variability in the number of site visits provided by each Organisation is shown in Figure 2.

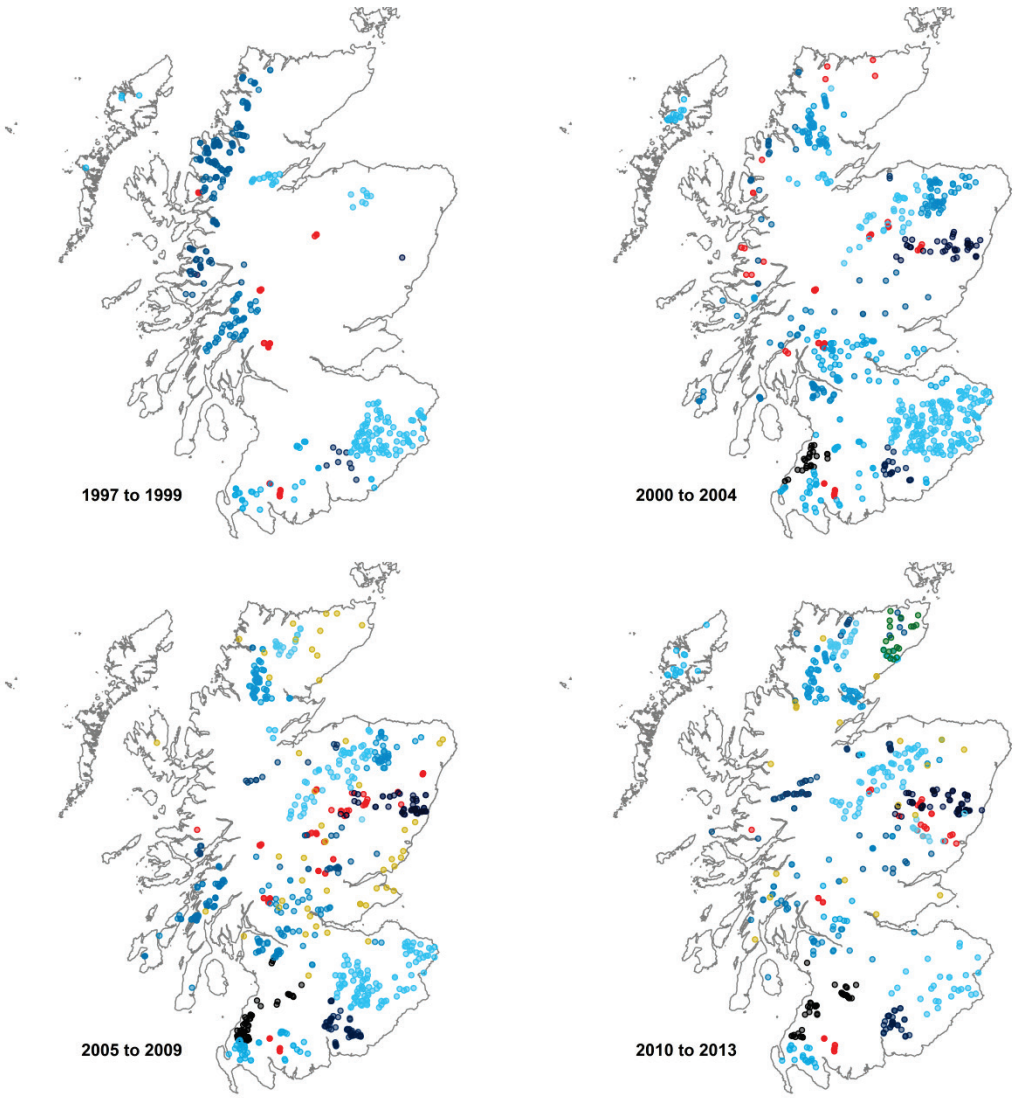

Figure 1 Spatial and temporal data coverage (un-stocked sites with multi-pass electrofishing, below impassable barriers) between 1997 and 2013. Prior to 1997 there are too few data from too constrained an area for useful large scale model fitting. Data are colour coded by source: MSS (red), SEPA (yellow), Other (green), SFCC (each trust is represented by a different shade of blue).

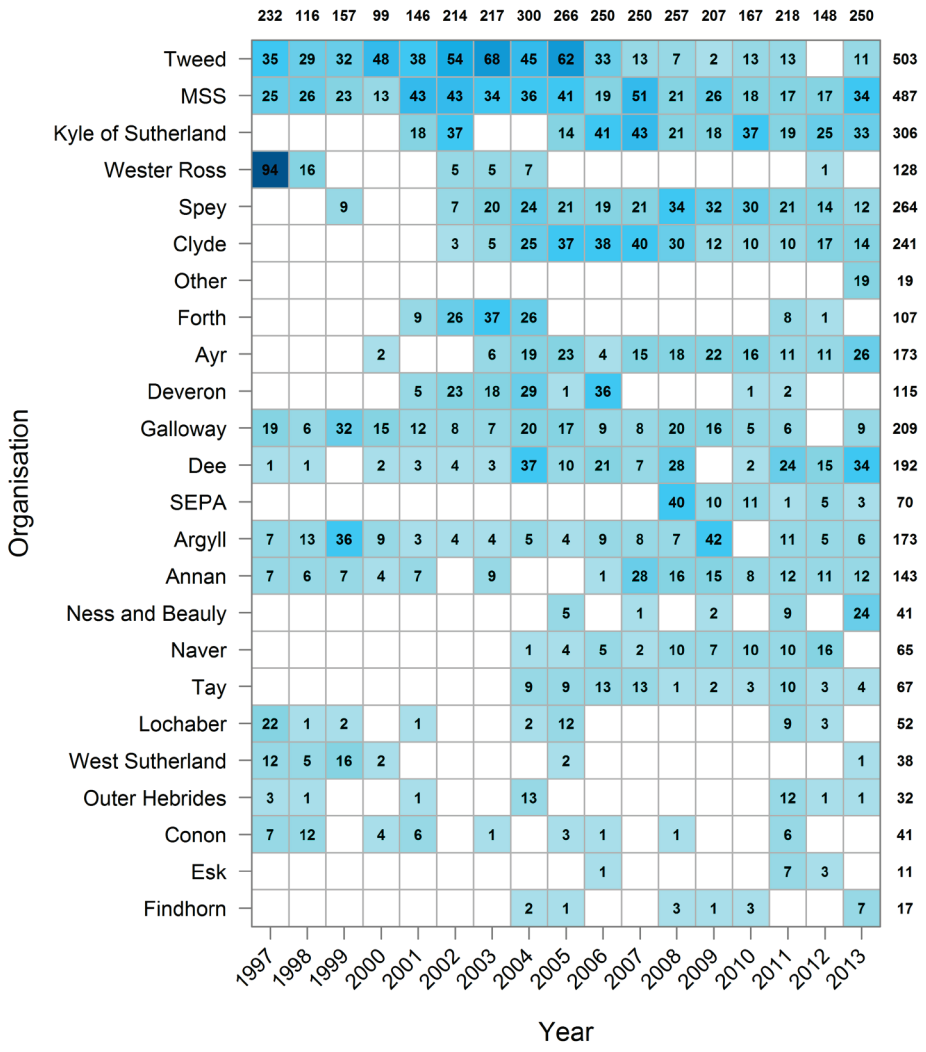

Figure 2 Schematic showing the temporal coverage of data by provider. Organisations are ordered by the mean annual number of site visits. Total site visits by Year and by Organisation are given in the margins.

Covariates

Covariates were obtained for each sampling location. The selection of covariates was informed by previous habitat modelling studies (Fausch, 1988; Niemela *et al.*, 2000; Wyatt 2005; SNIFFER, 2011) or relationships between covariates and processes that influence abundance, or capture probability. For example Altitude influences river temperature and in turn potentially affects fish productivity and capture probability, whereas Organisation may be expected to affect capture probability through differences in equipment or personnel, but should not affect fish productivity directly.

The continuous spatial covariates were: Upstream Catchment Area (UCA), Altitude, river network Distance to Sea (DS), Gradient, Channel Width (Width), landuse (see below for details). Data provider (Organisation) was included because of possible effects on catch probability. Hydrometric Area (HA) was included as a large scale spatial covariate. Catchment was also included to allow for finer scale deviation from regional trends. Options were also explored for including river network data that described spatial relationships between sampling points, hereafter referred to as River Connectivity (RC). Year and Day of Year (DoY) were included to allow for temporal variability. The location of sites above or below fish barriers was characterised to allow further data refinement depending on the species considered.

Before calculating spatial covariates it was necessary to assign (snap) sampling sites to a spatial dataset that represents Scotland's rivers. In this case the SEPA rivers dataset was chosen to provide the base mapping, where rivers are represented by a series of connected line features. In the case of MSS electrofishing data, a manual check of locational accuracy was performed, but this was not possible for externally sourced data. The resulting layer of snapped sites was used to obtain spatial covariates, the details of which are described below. All GIS analyses were performed in ESRI ArcGIS version 10 unless otherwise stated.

The selected covariates can be broadly assigned to four groups (1) spatial covariates (Hydrometric Area, Catchment and River Connectivity), (2) habitat covariates (Altitude, Upstream Catchment Area, Distance to Sea, Gradient, Landuse and Channel Width (3) sampling covariates (Organisation) (4) temporal covariates (Year and Day of Year).

1. Spatial Covariates

Hydrometric Area

SEPA hydrometric areas are administrative regions including a single large catchment or a number of contiguous smaller catchments with similar topographical characteristics (Marsh and Hannaford, 2008). They represent an intermediate spatial scale useful for modelling correlated regional variability in fish abundance. All Sites were attributed to a hydrometric area using the 'overlay' function in the R package rgdal and the SEPA hydrometric area polygons shapefile. In this report, HA was a regional identifier that links to additional information on connectedness to other regions (see 'Adding Regional Smoothers' in the Software section).

Catchment

All sampling points were allocated to a SEPA Baseline or Coastal river catchment. SEPA baseline rivers are only those catchments with an area of >10km^2. Catchments were too numerous for describing large scale regional variability in fish abundance, but were useful for describing finer scale deviations from regional trends.

River Connectivity

To describe spatial variation on a network, it is only necessary to know the connectedness of sampling locations and confluences. There is therefore considerable redundant information in the full river network dataset. The R package igraph (an R package for analysis of mathematical graphs), describes the connectedness of sampling points on a river network and the function 'reduceNetwork' reduces the information to a manageable size (see 'Adding a River Network Smoother' in the Software section). The River Connectivity covariate is essentially an identifier for a node on the river network that links to additional information on connectedness to other nodes.

2. Habitat Covariates

Altitude

A National Digital Terrain Model (DTM) provided by Centre for Ecology and Hydrology (CEH), was used to derive both altitude and slope (see below) for each sampling point. Altitude was obtained directly from the DTM using the 'Extract Values to Points' function.

Upstream Catchment Area

Upstream catchment area was obtained using the accumulation grid dataset provided by the CEH and the 'Raster Calculator' tools within spatial analyst. The latter calculates the number and size of upstream cells contributing to each cell in the CEH DTM. In common with the altitude and slope extractions, values for upstream areas were obtained using the 'Extract Values to Points' function from the resulting raster file.

Distance to Sea (along river network)

Rivers within the SEPA rivers dataset are split into connected line segments of known lengths and organised so that it is possible to take a point on the river and measure its route to the mouth. The distance to mouth was calculated for each sampling location using the 'Locate Features Along Routes' tool within 'Linear Referencing'.

Gradient

A slope raster was created within the 'Spatial Analyst' toolbox using the CEH DTM. For each 50m cell, the 'Slope' tool calculates the maximum difference in height between that cell and its surrounding neighbours and calculates the maximum gradient in degrees. Using the new raster, slope values were then obtained again using the 'Extract Values to Points' function.

Landuse

Metrics of land use and channel width were obtained from the Ordnance Survey MasterMap dataset. An automated script (within the 'select' option) was used to group polygons into land use themes, including inland water (rivers and lochs), Marsh, Urban, Mixed Woodland, Deciduous Woodland, Conifer Woodland and Other. The 'buffer' tool was then used to create 50m diameter circular polygons around each sampling location from which percentage landuse characteristics were obtained having removed the area associated with inland water polygons. For each theme, the 'intersect' command was used to ascertain the overlap with the site buffer. The 'dissolve' option, which aggregates features (i.e. unique site name) based on specified attributes (i.e., total area of specific land use type) was then used to merge the multiple theme polygons. The area for which riparian landuse was characterised varied depending on river width from a maximum of 1965m^2 to

a minimum of 85m^2. Only 2.5% of sites were characterised by land areas of <1000m^2. As such, although with limitations, this approach provided a useful metric of “local landuse” in the vicinity of the sampling location.

Channel Width

A similar approach to that used to determine landuse was used to approximate mean channel Width from the perimeter and area of inland water polygons within the buffer using the following equation:

$$0.25\ (Perimeter - \sqrt{(Perimeter^2 - 16\ Area)})$$

This approximation is increasingly accurate for polygons with a more rectangular shape, but nevertheless should identify major differences in channel width where this can vary over > 2 orders of magnitude. Unfortunately OS MasterMap defines all waterbodies less than 1m (urban areas) or 2m (rural areas) in width as line features. This results in areas and perimeters of 0. A manual correction was therefore applied to these cases with fixed values of 1.0 m where land use type was predominantly urban or 2.0 m elsewhere.

3. Sampling Covariates

Organisation

All data were assigned an organisation. There was some spatial structuring in the data from fisheries trusts. However, data provided by MSS and SEPA was geographically spread, and some trusts generated data in other trust areas thereby reducing the potential for confounding the effects of organisation and space (which was also represented by hydrometric area).

Sources of Error in Covariates

A number of technical problems were identified while generating spatial covariates. At the most basic level, sites often had incorrect grid references which placed them in the wrong catchments or even in the sea. Fortunately, these problems were straight forward to identify and the data could be omitted. Less obvious problems arose where site locations were recorded with low precision or lesser degrees of inaccuracy. Under such circumstances sites could be snapped to the wrong river line (Fig. 3) with serious consequences for the estimation of certain covariates. These issues were possible to correct

where the electrofishing locations were known, but not where data was provided by other organisations.

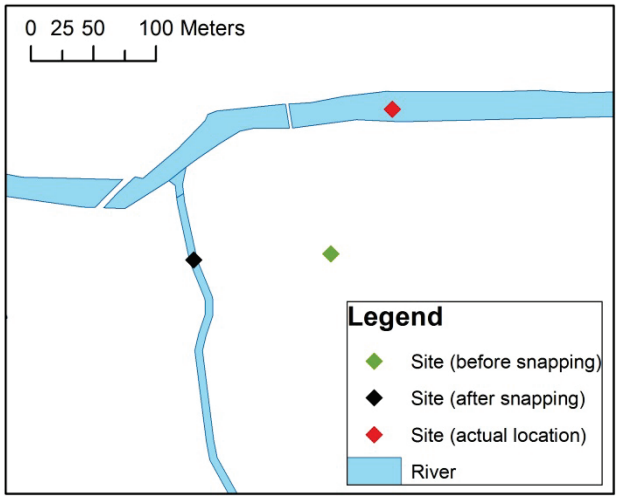

Figure 3 Example showing the consequences of low precision or inaccurate measurements of Site location. In this example the recorded site location was not on a river. Snapping to the nearest river places the Site on a tributary rather than the main-stem river where it was actually located. Based on digital spatial data licensed from the Centre for Ecology and Hydrology, © NERC. © Crown copyright. Licence 100024655.

Problems also arose where spatial data had been captured at different resolutions and were therefore not completely coherent. For example, the SEPA rivers line dataset (derived from the CEH rivers dataset) was not always spatially coherent with the OS MasterMap polygon dataset. This caused problems because it was necessary to snap site locations to the SEPA rivers, but then to estimate river widths from MasterMap. Where the two were not coherent this could result in erroneous estimates of channel width or other landuse characteristics (Fig. 4). Similarly, there were circumstances where sites could be incorrectly allocated to catchments because of the spatial resolution and precision of catchment boundaries (Fig. 5) or due to other unknown errors (Fig. 6). However, these issues were resolved by identifying the river on which points were located rather than the specific location.

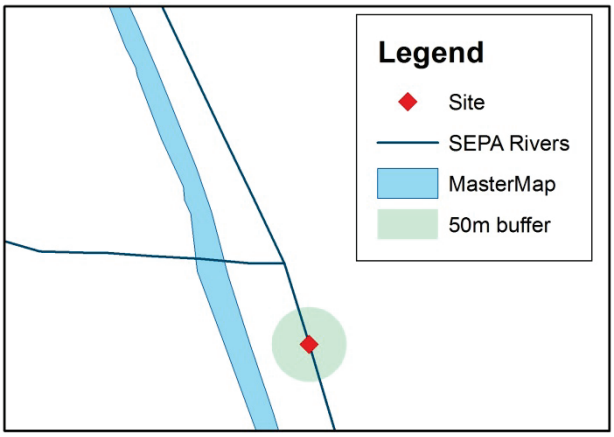

Figure 4 Example of circumstances where SEPA rivers line features and MasterMap polygons were spatially incoherent. In this example the buffer does not contain any inland waters polygons. As such this would result in an estimated mean channel width of zero. Based on digital spatial data licensed from the Centre for Ecology and Hydrology, © NERC. © Crown copyright. Licence 100024655.

Finally, there were circumstances where the SEPA rivers line dataset contained connectivity errors that needed to be corrected (e.g., Fig. 7). Each river segment in the dataset has a ‘from’ and a ‘to’ node which identify connections to other segments and flow direction. River sources have a ‘from’ node that does not reference to other valid nodes. Similarly river mouths have a ‘to’ node specified in the same way. Problems arise where the ‘to’ and ‘from’ nodes are misspecified so they are connected to distant rivers, flow in the wrong direction or appear as breaks in the network. Problems were identified by visually inspecting plots of the routes between sampling sites and river mouths. In practice, resolution of these problems was achieved by re-specification of all river node names in the SEPA rivers dataset using their unique latitude and longitude combination.

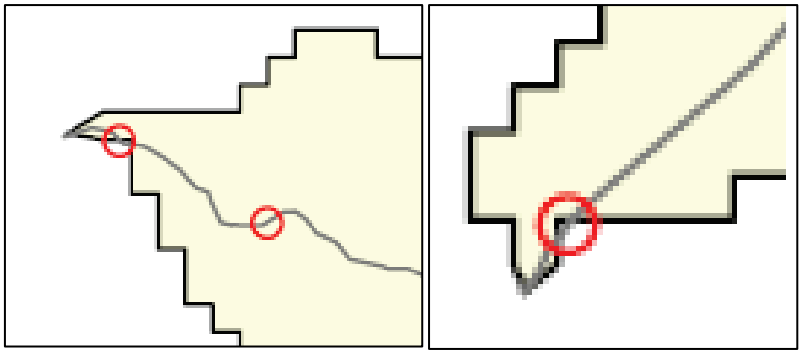

Figure 5 Examples of differences in spatial resolution resulting in incorrect allocation of Sites to Catchments. Red circles indicate sampling sites, black lines catchment boundaries and grey lines, rivers. Based on digital spatial data licensed from the Centre for Ecology and Hydrology, © NERC.

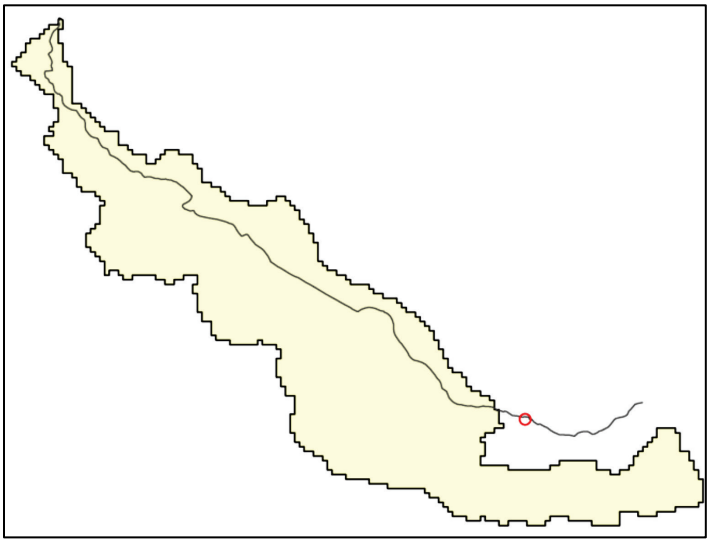

Figure 6 Examples where a SEPA river line runs outside of the catchment polygon. The sample point (red dot) lies on the river (grey line) but is outside the catchment boundary polygon (black line, shaded). Based on digital spatial data licensed from the Centre for Ecology and Hydrology, © NERC.

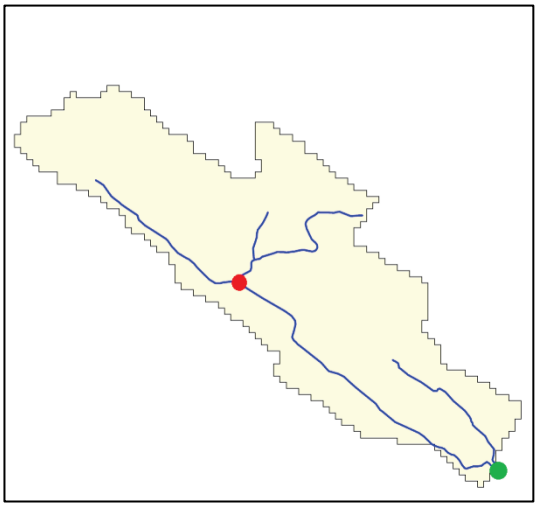

Figure 7 Example of a false river mouth due to un-connected river segments. The true river mouth (green circle) is in the south east, but there is an extra river mouth (red circle) assumed.

Model Development

Because a combined model of capture probability and fish density was not considered to be practical for a national scale model due to computational complexity, a two stage approach was developed here. Capture probability was modelled assuming independent estimates of densities for each site visit (i.e., a full density model). Spatio-temporal variability in density was then modelled based on these estimates of capture probability. The two main advantages of this modelling process are 1) model selection is simplified because only one aspect of the model (capture probability or spatio-temporal variability in density) is being tested at a time and 2) conditional modelling of density based on previously estimated capture probabilities allows the use of generalised additive models (GAMs) and structural additive regression (STAR) models (Fahrmeir *et al.*, 2013). This allows the incorporation of a wide range of penalised smoothers and spatial and random effects as well as the option of using a negative binomial distribution over the Poisson to allow for additional Site Visit variation.

Development of a Two Stage Model: Model Theory

The joint likelihood for capture probability (p) and fish density (λ) can be defined as:

$$L(p,\lambda;X,T) = p^{T}(1-p)^{X}(\text{Area}\,\lambda)^{T}\exp\left[-\left(1-(1-p)^{S}\right)\text{Area}\,\lambda\right] \quad (1)$$

where, $T = n_1 + n_2 + \ldots + n_S$, and $X = n_2 + 2n_3 + \ldots + (S-1)n_S$ and n_i (i = 1, …, S) is the number of fish caught on the ith of a total of S passes and Area is the area fished.

Stage 1: Capture probability model

For any given capture probability the estimated fish density for a site visit can be defined as:

$$\hat{\lambda} = \frac{T}{\left(1-(1-p)^{S}\right)Area}. \quad (2)$$

Because the estimate of λ can be written in terms of p (Eqn. 2), it is possible to define a likelihood for p in terms of $Z = X/T$ and T as follows:

$$\log L(p; Z, T) = T\,(\log p + Z \log(1-p) - \log(1-(1-p)^{S})). \quad (3)$$

Capture probability can then be modelled in relation to covariates using the logistic link. The likelihood defined in (Eqn. 3) can then be used in model selection.

Stage 2: Fish density model

Given an estimate for capture probability, p, the distribution of T is Poisson with expectation:

$$E[T] = \lambda Area\left(1-(1-p)^{S}\right) \text{ fish}, \quad (4)$$

Given estimates of capture probabilities from Stage 1, densities can be modelled on the log link using GAMs and incorporating an offset of log *Area* + log $(1-(1-p)^{s})$. Because electrofishing data are often more variable than assumed by a Poisson distribution a negative binomial distribution can be used to account for over dispersion.

Covariates and Model Selection

Both capture probability and densities can be modelled in relation to categorical, linear or smoothed covariates. While the incorporation of standard smoothers or those describing regional spatial variability are possible in mgcv, River Connectivity cannot be fitted in the standard package. This may be important in situations where sampling points are geographically similar, but are far apart in terms of river network distance. Similarly, there are no existing packages that can allow covariates to be added to capture probability models (above). These requirements were considered during this project where a new package was developed to combine and extend existing functions in R, particularly from the packages mgcv and igraph and to a lesser extent the spatial package maptools, rgdal, rgeos.

In the case of capture probability, the model building functionality of mgcv (Wood, 2011) was used to generate a design matrix to include covariates in the likelihood from (Eqn. 3) to allow rapid specification of complicated models. However, this method does not currently allow the use of generalised cross-validation to determine the degree of smoothing and as such is necessary to specify fixed degrees of freedom.

One of the main benefits of the two stage modelling approach was the ability to rapidly specify, fit and compare between models In the case of the capture probability model this can be achieved using BIC (Schwarz, 1978). However, in the case of the density model (which is a GAM), it is possible to use more sophisticated (ridge regression) model selection from the MGCV package to automatically remove terms.

Software: Data Preparation Requirements and Model Specification

The models and model fitting were provisionally implemented in an R package ‘CLModel’. The package is currently only available within MSS pending testing and peer review of the methods. The package has the following dependencies which are available from CRAN: mgcv, igraph, maptools, rgdal, rgeos, sp, spdep. In addition, the library rstan is required and this is obtained from its own dedicated website (http://mc-stan.org/rstan.html).

Data Preparation

Data should be prepared in a data-frame with covariates stored in columns and lines representing site visits. The package is flexible in terms of column headers, but the data-frame must contain the number of fish caught on each pass (separate columns) and the number of passes. Where there is an incomplete set of covariates these data are excluded from any model fitting, although this was not a concern in the current exercise. The function ‘getData’ must be run on the data-frame to provide electrofishing summary statistics before any further analysis. Following estimation of capture probabilities, it is necessary to calculate an offset before modelling fish densities. This is done using the function ‘getOffset’ or manually (see ‘Development of a Two Stage Model’ above). If calculated manually this should be stored in a column labelled ‘offset’.

In the following examples the data is assumed to reside in a data-frame called ‘Electrofishing’.

Adding Factors, Linear Effects and Smoothers

Factors (e.g., Organisation or HA), linear effects (e.g., Width) and smoothers (e.g., DoY) can be specified for the capture probability model as follows:

```
efp(Z ~ Organisaton + Width + s(DoY, k=3), data = Electrofishing)
```

where in this example DoY is restricted to a smoother with 3 degrees of freedom. Similarly in the case of the fish density model:

```
gam(T ~ HA + Width + s(DoY, k=6),
    offset = offset, data = Electrofishing)
```

where this time, because the flexibility of the smoother is estimated from the data, a larger (maximum) degrees of freedom can be safely specified. Other standard GAM functionality remains. A range of ‘smooth types’ are possible e.g., thin plate splines with shrinkage (bs=‘ts’) or random effects (bs=‘re’) as in

```
gam(T ~ HA + s(Catchment, bs=‘re’) + Width + s(DoY),
    offset = offset, data = Electrofishing)
```

where HA is modelled as a factor and catchment as a random effect.

Adding Regional Smoothers

Regional covariates such as HA could be estimated independently (i.e. as a factor), but given spatial correlation it is more appropriate to fit a model where neighbouring regions have similar effects. This requires specification of the spatial (neighbourhood) structure using poly2nb and nb2mat functions in the spdep package. The shape file should be read into R using readOGR from the rgdal package and stored in an object (in this case 'ha') where the neighbourhood connections are calculated and converted to an adjacency matrix before conversion to GMRF using the function 'getRegionalGMRF' from the CLModel package.

```
ha_adj <- poly2nb(ha, queen = FALSE)
ha_adj <- nb2mat(ha_adj, style = 'B', zero.policy = TRUE)

Qha <- getRegionalGMRF(ha_adj)
```

The regional effect for HA is then specified using bs='gmrf', a smoother type extending mgcv functionality developed for the CLModel package.

```
gam(T ~ s(HA, bs='gmrf', xt=list(penalty=Qha)) +
        s(Catchment, bs='re') + Width + s(DoY),
    offset = offset, data = Electrofishing)
```

Although it is possible to incorporate spatial smoothers using existing functionality in mgcv, the 'gmrf' extension provides a more flexible interface that can be used to fit both regional and network scale effects. It can also automatically accommodate regions without observations but where there is information on spatial structure. By default the maximum degrees of freedom for regional smoothers would be equal to the number of regions. However, this can be restricted using by specifying a lower value e.g., 'k=10'. If a regional smoother were to be included in the capture probability model, restricted degrees of freedom should be specified.

Adding A River Network Smoother (River Connectivity)

Although a River Connectivity model can be applied to single catchments, it has not yet been possible to extend this functionality across multiple catchments. Consequently, in the following, the data-frame 'Electrofishing' is assumed to contain information from a single catchment. River Connectivity (see Covariates section above) is stored as a graph object and calculated

using the ‘buildTopo’ and ‘reduceNetwork’ functions in the CLModel package incorporating functionality from the rgeos and igraph packages. The function ‘getWRW1Mat’ in the CLModel package converts the graph object into a GMRF.

```
graph <- buildTopo(rivs)
graph <- reduceNetwork(graph)
Qrc <- getWRW1Mat(graph)
```

The ‘getRCLocation’ in the CLModel package is then used to identify the location of sites on the network and to link to the GMRF (i.e., generate River Connectivity covariate).

```
Electrofishing $ RC <- getRCLocation(Electrofishing, graph)
```

A model containing River Connectivity can then be specified as follows:

```
gam(T ~ s(RC, bs=‘gmrf’, xt=list(penalty = Qrc)) +
        Width + s(DoY),
    offset = offset, data = Electrofishing)
```

Predicting

It is straightforward to predict fish densities on a river network smoother using standard methods in mgcv. In the following simple example, the densities are predicted only on the basis of RC and the model fit is stored as an object ‘m’. Predictions are then made across the entire network by:

```
m <- gam(T ~ s(RC, bs=‘gmrf’, k=12, xt=list(penalty = Qrc)),
         offset = offset, data = Electrofishing)
pred <- predict(m, newdata = data.frame(RC = V(graph)$name))
```

In the case above the new data-frame used for prediction consists of a single column ‘RC’, containing the id’s of all the nodes on the entire river (sites and confluences).

Model Fitting: A Case Study Using 0+ Atlantic Salmon

To demonstrate the modelling approaches described in this report an analysis of salmon fry abundance was undertaken. All potential covariates (see above) were included during model selection for both capture probability and density with the exception of Organisation, which was only included in the capture probability model because it should not affect fish density; and catchment that was only included in the density model because there were not electrofishing observations in all catchments resulting in over fitting. Given the large number of potential interaction terms only main effects were considered.

Data coverage

Pairwise density plots of the data coverage in relation to covariates are presented in Figure 8. Latitude and Longitude have been included to identify the spatial coverage of available data, although they were not included as covariates during model fitting. The plots highlight where combinations of covariate values are well represented, for example low slopes and short distance to sea. But also where combinations are absent or rare, for example low elevation, high distance to sea. In addition, some combinations of land use variables are not possible because the landuse combinations need to sum to 100%.

Capture probability

The final capture probability model was selected using a step-wise selection procedure based on minimum BIC, where the initial model had a common capture probability across all site visits. Organisation and Year were included as factors. Year was included as factor to allow for inter-annual variability in hydrological conditions and on the assumption that there would be no reason to assume long-term trends in capture probability. HA was included as a spatial effect. All remaining continuous variables were included as both linear effects and as smoothers with 2 degrees of freedom. The final model was:

$$\text{logit } p \sim \text{Organisation} + s(\text{DoY}) + \text{Year} + \text{regional(HA)} + \text{DS} + \text{Width} + s(\text{Gradient})$$

The relative importance of explanatory covariates was assessed by dropping terms one at a time from the final model. The importance of terms in the final model was indicated by the magnitude of changes in BIC with more important terms associated with greater changes (Table 1). Organisation was the most

important covariate, followed by DoY, Year, DS, Width and Gradient. The addition of Gradient has a minimal effect on the fit.

Covariate	Degrees of Freedom	Change in BIC
Organisation	23	889.2
s(DoY)	2	438.1
Year	16	207.4
regional(HA)	7	176.2
DS	1	57.3
Width	1	27.7
s(Gradient)	2	0.6

Table 1 The relative importance of explanatory covariates in the capture probability model as indicated by changes in BIC where single terms were removed from the final model. The degrees of freedom shows the reduction in parameters associated with removing the term.

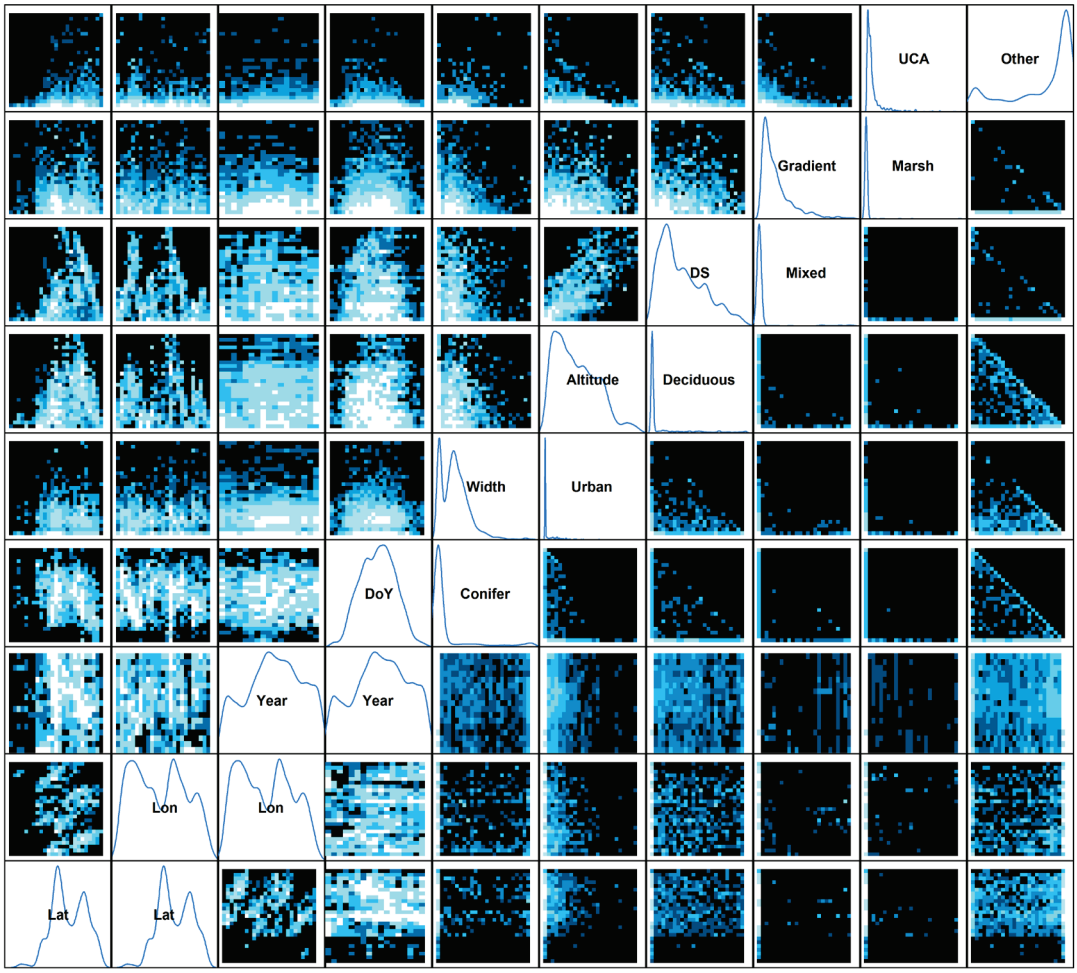

Figure 8 Density plots showing the distribution of available data in relation to combinations of environmental covariates (white: lots of data, blue: few data, black no data). Latitude (Lat) and Longitude (Lon) are included to provide an indication of spatial coverage although these were not included in model fitting.

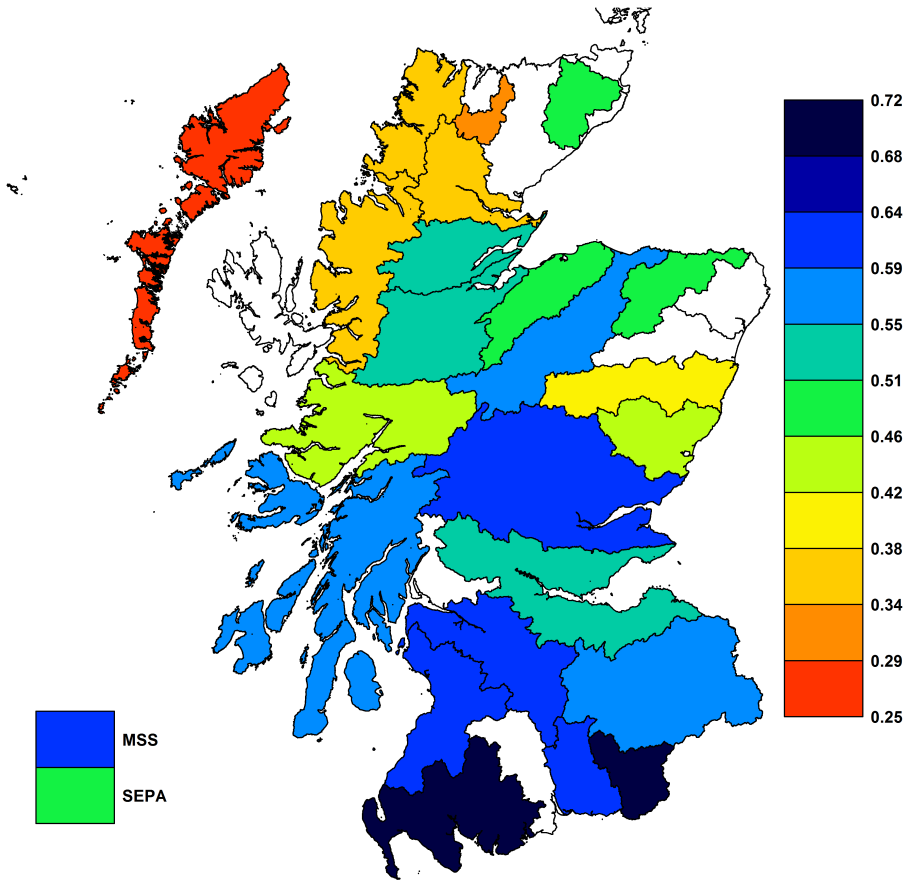

Figure 9 Estimates of capture probability by data provider (Organisation). Estimates are plotted in relation to the geographic area of responsibility (e.g., Trust Boundaries) of each organisation, although data coverage may be broader. Note that the Galloway Fisheries Trust covers the Galloway region and the Border Esk to the East. White areas indicate either no multi-pass fishing data or no data provider present in that region. In the case of the River Don, permission to use available data came too late to include in this report. MSS and SEPA estimates are indicated to the side of the map given their wide ranging data coverage. Estimates are conditioned on HA Tay, Year 1996 and median values for all remaining covariates. Missing estimates reflect a lack of data availability at the time of model fitting. Map based on digital spatial data licensed from Centre for Ecology and Hydrology, © NERC.

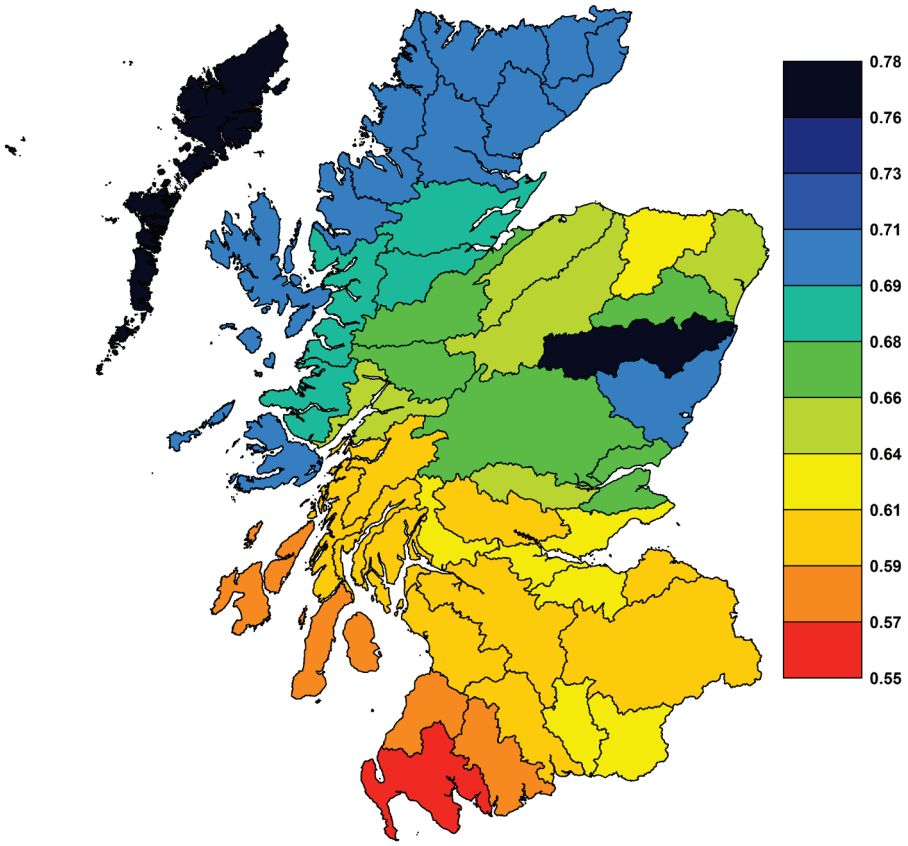

Figure 10 Estimates of spatial variability in capture probability (HA). Estimates are conditioned on Year 1996, Organisation MSS, and median values for all remaining covariates. Map based on digital spatial data licensed from Centre for Ecology and Hydrology, © NERC.

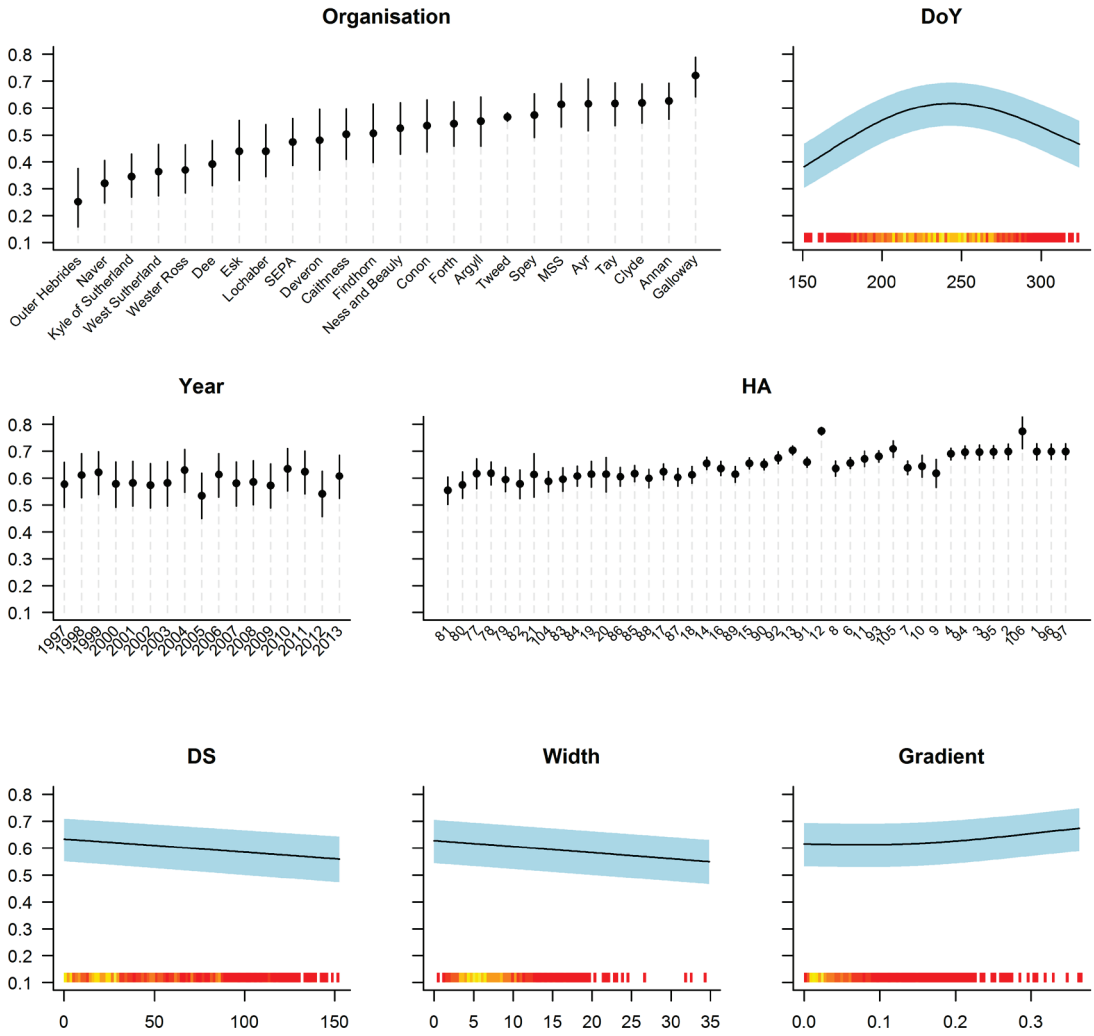

Figure 11 Relationships between capture probability and covariates. Plots are conditioned on HA Tay, Organisation MSS, Year 1996, and median values for remaining covariates. Organisation names have been abbreviated. HA values are ordered from South to North. 95% pointwise confidence intervals are shown as shaded blue areas or vertical bars. A ‘rug’ indicates the distribution of available data on the x-axis (red: few values, yellow: many values).

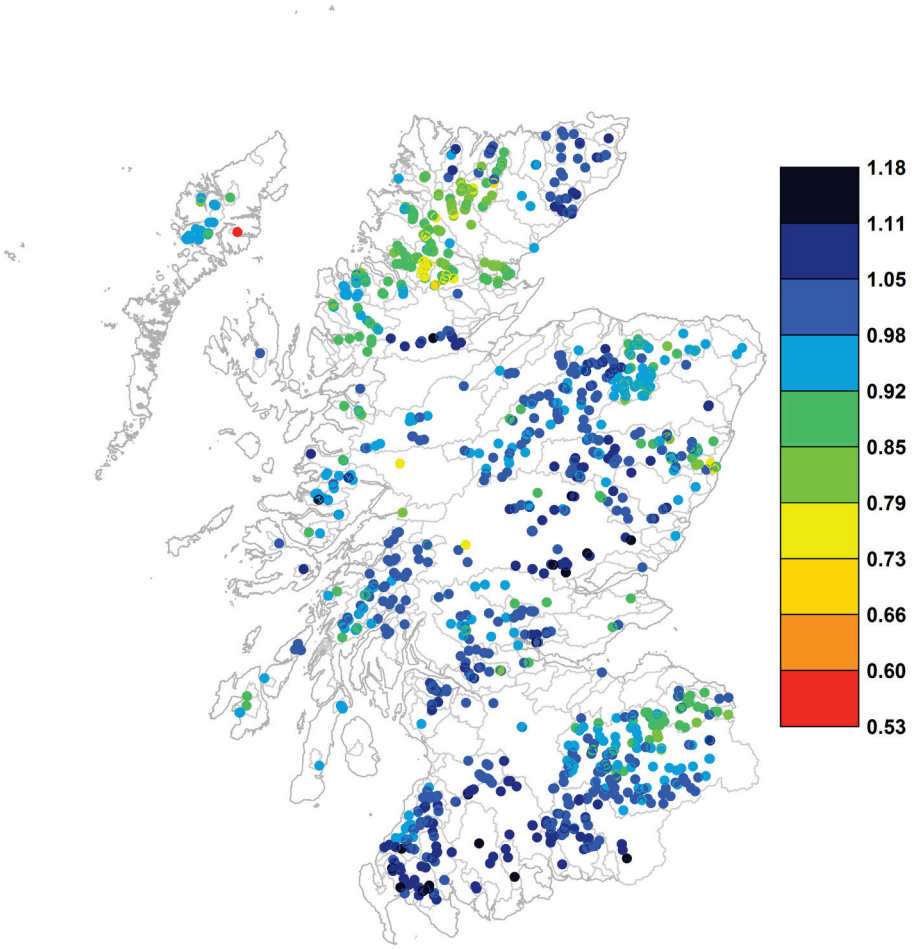

Figure 12 Map showing proportional differences in density estimated using a constant mean capture probability (0.53) and modelled capture probabilities. Higher values indicate higher modelled densities relative to constant capture probability. Map based on digital spatial data licensed from Centre for Ecology and Hydrology, © NERC.

Organisation had the greatest effect on capture probability, potentially reflecting differences in equipment and procedures between data providers (Fig. 9 & 11). The next strongest effect related to the time of sampling (DoY). This had a modal effect of similar magnitude which increased from day 150 to ca. 240 and declined thereafter (Fig. 11). Year had a substantial effect, but showed no temporal trends. HA also had an effect of similar magnitude to year, but also showed an increasing trend in capture probability from south to north (Fig. 10 & 11). The remaining effects were relatively small. Catch

probability showed a negative linear response to DS and Width and a positive non-linear response to Gradient.

Figure 12 demonstrates the importance of including catch probability in estimates of fish density. Densities were estimated for each site assuming 1) modelled capture probabilities and 2) a constant capture probability of 0.53 (the mean of the modelled capture probabilities). The plotted differences indicate where modelled capture probabilities result in proportionately greater (>1) or lower (<1) estimates of density. Clear spatial patterns can be observed in the resulting difference plot indicating that a failure to account for variation in capture probability would result in biased estimates of density and potentially misleading spatio-temporal models.

Density

Because the density model was fitted as a GAM it was possible to use ridge regression to drop terms in a form of automatic model selection (Wood, 2006). Models were fitted using Restricted Maximum Likelihood (REML) assuming a negative binomial distribution to allow for over-dispersion. Any retained terms where the degrees of freedom were estimated to be one or less were replaced by linear terms. The number of degrees of freedom was restricted to a maximum of 3 for all continuous variables. The HA spatial smoother was restricted to a maximum 24 degrees of freedom. The final model was:

log density ~ random(Catchment) + random(Year) +
s(DoY) + s(Width) + s(Altitude) + s(DS) +
Gradient + s(UCA) + Urban + Conifer +
s(Mixed) + s(Other)

Catchment and Year had substantial spatial and temporal effects on fish density respectively, although there was no evidence of a long-term trend in the data i.e. year was retained as a random effect, but not a temporal smoother (Fig. 13 & 14). The greatest “habitat effects” were associated with DS, UCA and Altitude, which exhibited asymptotic and, complex non-linear and decreasing non-linear responses respectively (Fig. 13). DoY and Width were the next greatest effects. The effect of river Width was modal around 8m, but was highly uncertain above ca. 25m, where there were very few data, and observations may reflect difficulties in assigning appropriate Widths or biased partial sampling of wider rivers (see section ‘Sources of Error in Covariates’). Density estimates exhibited a negative non-linear trend with

DoY. The remaining effects (Gradient, % Mixed Woodland, % Conifer Woodland, % Urban and % Other landuses) were substantially smaller and negative with the exception of Other. Percentage Marsh, Deciduous woodland and HA were dropped from the final model.

The fish density model predicted spatially heterogeneous fish densities, although the highest densities were usually predicted for large east coast rivers, with lower densities generally in the central belt and west coast (see Figure 14: Fitted Values). To improve understanding of the effects of individual covariates on the spatial distribution of densities, their conditional effects were explored (Fig. 14).

The effects of catchment were highly heterogeneous with nearby catchments often exhibiting strongly contrasting effects e.g., the Nith and Kirkubrightshire Dee. However, catchment effects were generally negative in the central belt reducing density expectations in these areas.

The effect of DS was to increase densities in inland areas of larger east coast catchments (Fig. 14: DS). In contrast, low altitude areas were predicted to have higher densities and these were observed near the coast of all rivers (Fig. 14: Altitude). The effect of UCA was to mirror DS, increasing predicted densities in larger rivers close to the sea, but also favouring overall densities in large dendritic rivers such as the Tweed. Given the spatial correlations between DS, Altitude and UCA, it is useful to see the combined effects of these variables which reduced predicted densities in major upland areas such as the Cairngorms and predicted higher densities in lowland and coastal areas (Fig. 14: DS, UCA, Altitude). Given the modal effect of width, the spatial distribution of effects was highly heterogeneous (Fig. 14: Width).

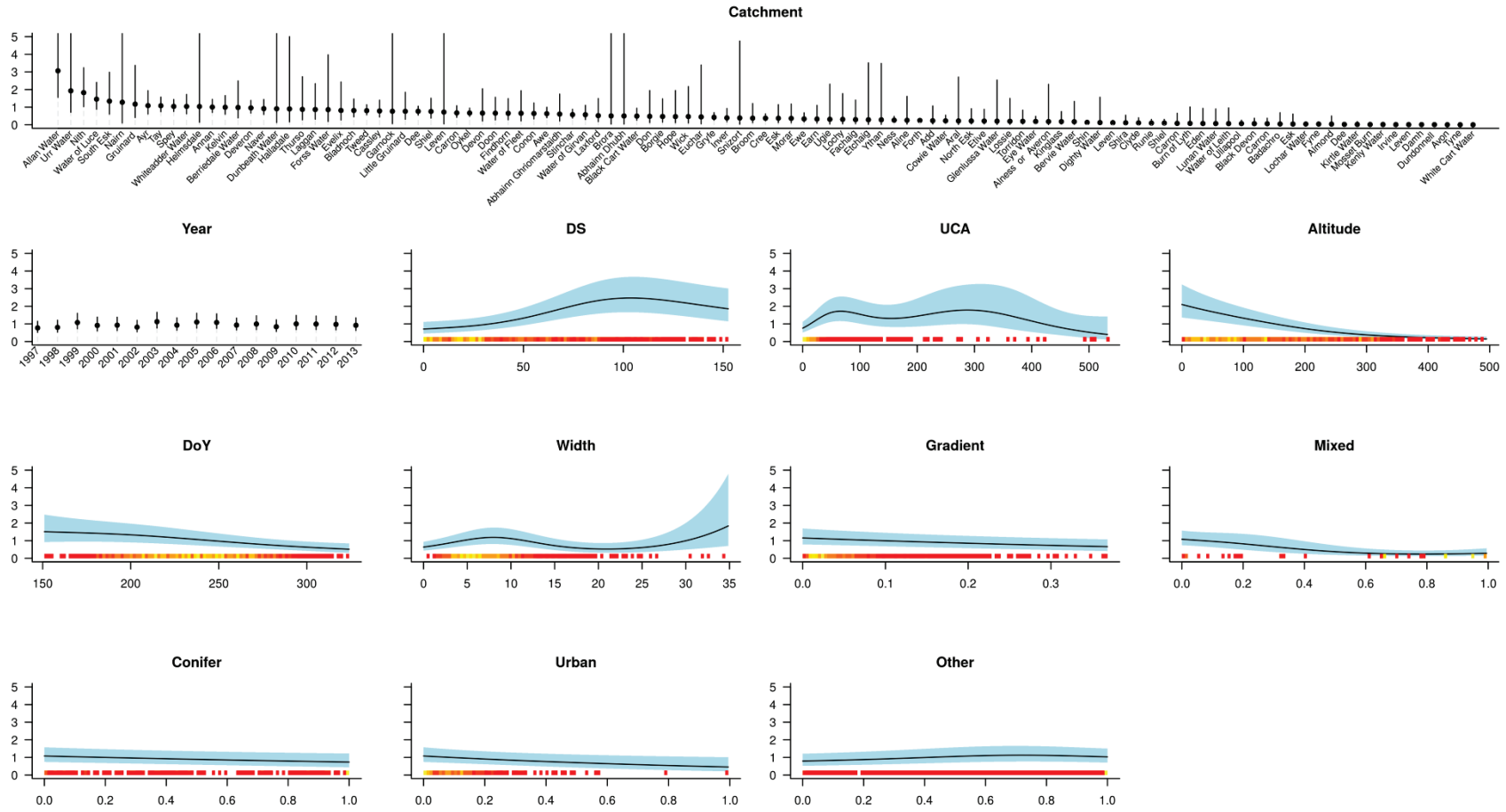

Figure 13 Relationships between fish density and covariates. Plots are conditioned on Catchment Tay, Year 1996, and median value for all remaining covariates. Catchment names are abbreviated. 95% pointwise confidence intervals are shown shaded in blue. For continuous covariates a 'rug' showing the distribution of available data is overlaid on the x-axis (red: few data, yellow: much data, white: no data). High observations for % Urban area represent groups of observations with similar values rather than single observations.

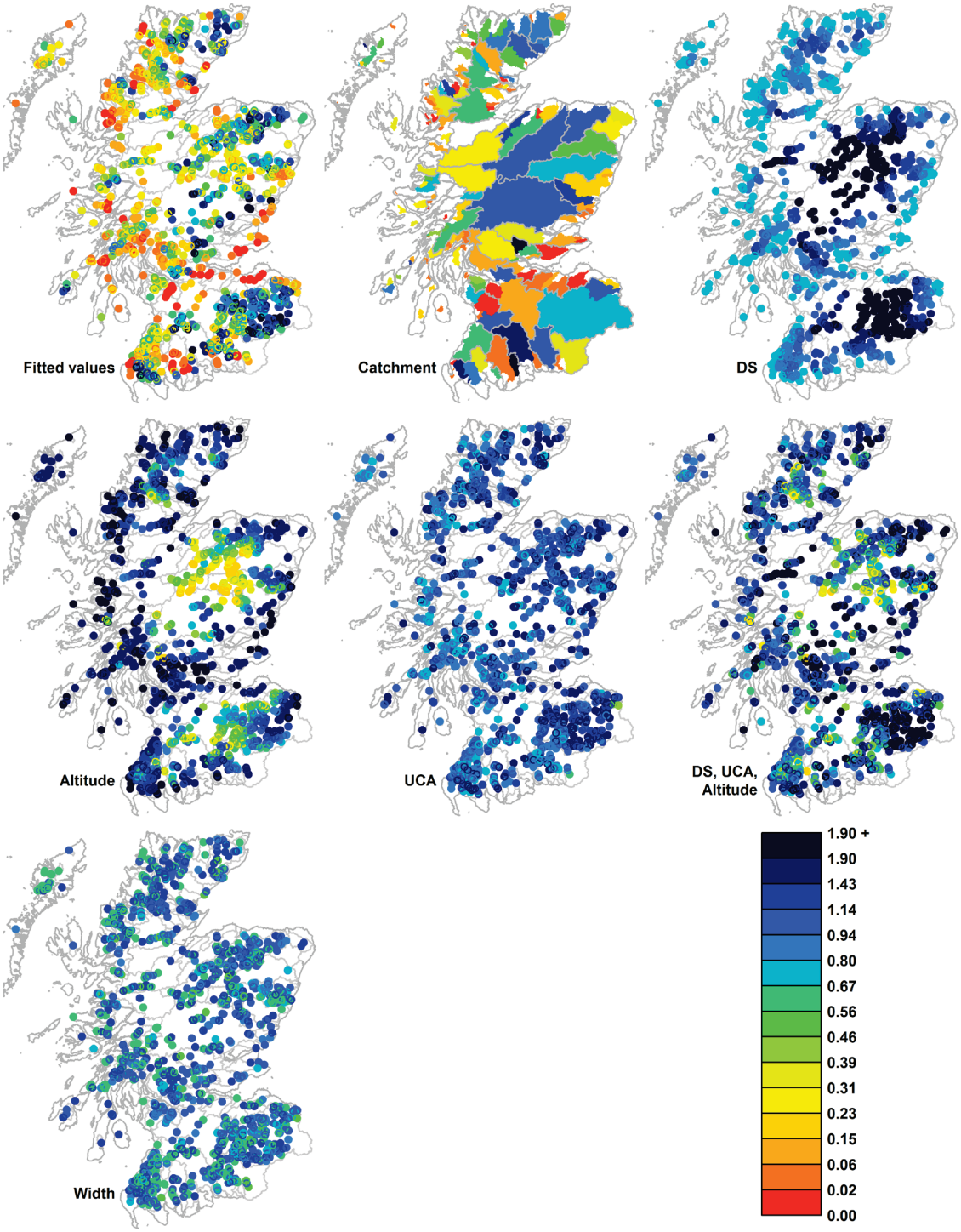

Figure 14 Maps showing the modelled effects of covariates (individual covariates and in-combination) on the spatial distribution of salmon fry densities. Model predictions are conditioned on Catchment Tay, Year 1996, and the median value for remaining covariates. Map based on digital spatial data licensed from Centre for Ecology and Hydrology, © NERC.

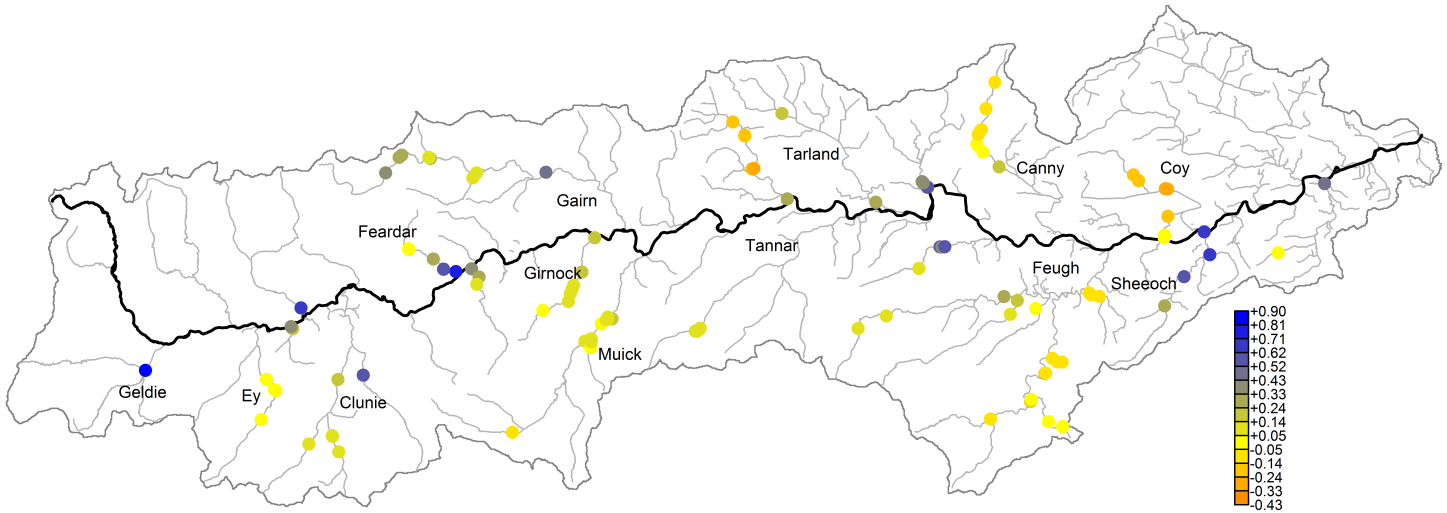

Figure 15 Performance of River Dee monitoring sites relative to the national expectation averaged across years. Coloured points indicate the difference between the predicted salmon fry densities from the national model (accounting for variability in habitat) and smoothed model residuals (fitted using the RC) including the effect of Catchment. Yellow points indicate that sites meet an average national expectation over the monitoring period, orange indicate lower than expected, green and blue higher than expected. Light lines indicate tributaries, the bold line indicates the mainstem river as defined by the SEPA river lines dataset. Map based on digital spatial data licensed from Centre for Ecology and Hydrology, © NERC.

Having accounted for large scale spatial, temporal and habitat covariates, within river variability can remain in the form of network spatial correlation. This variation can be described using the RC smoother. Although it wasn't possible to fit this term to all rivers in Scotland, it was possible to demonstrate the effect for a single river catchment (the Aberdeenshire Dee). This was achieved by fitting RC to the density residuals for the Dee which results in a smoothed residual representation of the average (site-wise) deviations from model predictions (Fig. 15). Put more simply, the river smoother represents the within catchment deviation from the average national habitat model averaged over time. When combined with the catchment effect, the resulting predictions indicate the differences in density between average model predictions for Scotland and those for the Dee. Figure 15 suggests that salmon fry densities in the Dee are generally as expected (yellow) or better (yellow-green to blue) than would be expected from the mean national model with no marked spatial patterns, although poorly performing sites (orange) were only present in the lower half of the catchment.

Although RC was illustrated here using an average performance metric, it would also be possible to make predictions for the best year nationally (highest densities) and then investigate model residuals for each monitoring year individually. This would assess performance of monitoring sites relative to the best observed conditions and permit assessment of spatial variability in performance over time.

Discussion

There are a number of challenges involved in the development of large scale habitat models for predicting juvenile fish densities. These include estimation of densities from electrofishing data and consideration of spatial coherence at regional and network scales. Previous attempts to develop habitat models have generally followed one of three approaches (1) simplified approaches considering only single pass electrofishing data (Godfrey, 2005; Wyatt, 2005) or assuming constant capture probability (SNIFFER, 2011) (2) approaches involving site specific estimates of capture probability (Lanka *et al.,* 1987; Rosenfeld *et al.*, 2000; Godfrey, 2005) or (3) Hierarchical Bayesian approaches where capture probability and habitat covariates are considered simultaneously (Wyatt 2002, 2003; Rivot *et al.*, 2008). The first approach is clearly misleading since habitat variability in capture probability and abundance could be confounded. The

second approach does not make best use of available information, particularly in circumstances where no fish are caught and thus no estimate of capture probability can be obtained (Godfrey, 2005). The final approach often results in models which are time consuming to specify and fit, thus making model comparison and selection difficult.

In this report, a fourth approach was considered where relationships between habitat (characterised at large spatial scales using GIS covariates) and fish abundance were modelled in two stages. Firstly, capture probability was estimated in relation to covariates. Secondly, spatio-temporal variability in density was modelled as a function of fish numbers and covariates, conditioned on estimates of capture probability. This two stage modelling approach and the resulting R package, allowed improved estimates of capture probability, rapid specification, fitting and selection of models. Moreover, the R package allowed fitting of flexible (non-linear) relationships between fish abundance and habitat and for modelling spatially correlated data. Such advanced statistical approaches were not available to investigators for previous analyses of Scottish juvenile salmonid data (Godfrey, 2005).

Many large-scale models of salmon abundance have implicitly ignored the effect of capture probability, focussing on single pass data or data from the first pass of multi-pass fishings (Wyatt, 2005), or assumed a constant capture probability. This includes previous models of Scottish fish densities (SNIFFER, 2011). These approaches are inconsistent with the findings of a number of studies which have shown that capture probability varies with a range of factors including species and life stage (Borgstrom and Skaala, 1993), sampling protocol, personnel (Niemela *et al.*, 2000) and habitat characteristics (Kennedy and Strange, 1981). This report demonstrated that implicit (use of 1 pass data) or explicit (a single estimated capture probability) assumptions of constant capture probability are inappropriate simplifications. Capture probability varied substantially with organisation (which has a geographic component), with year, with the time of electrofishing and to a lesser degree with other habitat covariates. Failure to correct for differences in capture probability generates spatial and temporal biases and consequently misleading models of fish density.

Models of salmon fry density were successfully fitted to the available electrofishing data. Although some of the covariates (e.g., Channel Width) could

have been affected by data processing errors (particularly snapping to river), there is no reason to assume that these estimates should be biased and as such, these errors are only likely to introduce noise into the observed relationships. Furthermore, many of the important covariates which had a strong effect on abundance are robust to site allocation e.g., Catchment, Altitude, Distance to Sea. A potentially more serious issue involves the introduction of bias associated with partial sampling (single channel threads or edges) of large rivers. A number of these site visits were manually identified in the current project through inspection of individual data records. However, further clarification is required from data providers as to where this has occurred. Fortunately the issue of unrepresentative (potentially biased) sampling appears restricted to large and deep rivers where standard SFCC fishing protocols no longer apply. It is therefore likely that the models presented here provide a reasonable overall description of the spatio-temporal variability in fry abundance and the relative importance of covariates, although these preliminary results should be viewed with caution for sites with a high channel width (Mastermap river widths of ca. 27m).

Wyatt (2005) observed a modal response of 0+ salmon densities to Altitude and UCA about 155m and 123km^2 respectively. In this study we observed a non-linear negative relationship with altitude and a non-linear relationship with UCA. Although these results initially appear to contrast, the combined effect of DS (not illustrated in Wyatt, 2005), Altitude and UCA may produce similar modal spatial predictions.

Having accounted for between catchment differences in the distribution of habitat covariates (e.g., DS, Altitude) the Catchment spatial covariate describes residual variation in the abundance data. This could be considered an average relative catchment performance metric over the monitoring period, so long as observed differences did not reflect uncharacterised natural variability in habitat that affects productivity and sampling was representative of the catchment as a whole. Similarly, the RC covariate indicates within catchment variation in abundance relative to the catchment mean. The combination of Catchment and RC therefore provides the opportunity to examine the performance of areas relative to a mean national expectation.

The concept of reference or optimal condition underpins legislation such as the EU Water Framework Directive, while density expectations are required to interpret electrofishing data and fish population health for local fisheries management. Using the current model a reference condition could be inferred by selecting the best Year, setting Catchment to a median value and setting the effects of pressures (% Urban and potentially % Conifer if this represents commercial forestry) to zero. An alternative approach would involve modelling electrofishing data obtained from strategic stocking of sites over a broad geographic and environmental range or defining reference condition for a time period where there were high adult salmon returns and lower human impact, excluding sites thought to be affected by significant anthropogenic pressure. The latter approach would require historical data from relatively un-impacted (reference) sites. While the consequences of this study are clear in a number of contexts, further work would be required to consider how reference conditions could be used to inform any future development of conservation limits (e.g., Wyatt and Barnard, 1997).

If the models presented here were to be used to assess the status of juvenile salmonid stocks on an annual basis it would be necessary to establish a strategic national electrofishing program that combined quantitative multi-pass electrofishing with single pass area delimited electrofishing to maximise spatial coverage and minimise uncertainty in the resulting data. The optimal mix of quantitative and single pass fishing would require further consideration, as would the spatial structure of annual census data.

Recommendations for Future Work

Models are only as good as the data that underpins them. This report identified a number of problems with the location data and the generation of covariates. The location of sites was often incorrect or too imprecise. This caused particular problems where grid references placed sites near both tributary and mainstem rivers. This could be resolved by providing maps of ‘snapped’ electrofishing sites back to data providers and asking them to confirm the precise location of sampling which could improve model fits. There were also problems where the SEPA river line dataset did not match with the OS MasterMap dataset due to differences in the spatial scale of data capture. This resulted in erroneous

estimates of channel width. It is possible that this could be resolved by snapping sites to the forthcoming OS rivers dataset or by introducing rule based quality controls on the distance that points could be snapped to rivers and the adjacency of different rivers. These options would require further investigation. It should also be considered whether the size of the buffers could be adjusted depending on channel width, such that standard “land areas” could be characterised and whether additional landuse datasets could provide improved information on agricultural landuse.

In addition to the issues of site location and covariates, there were a number of occasions where it appears that partial and unrepresentative electrofishing surveys were performed on larger rivers e.g. sampling a single braid or sampling only the peripheral areas. In the context of the current analysis these partial samples could introduce bias. Again this could be resolved by asking data providers to identify any site visits where electrofishing did not encompass the full channel and excluding these data from the final analysis.

Despite successful model application, a number of developments could potentially improve the ability of these models to characterise and predict spatio-temporal variability in fish abundance at the national scale. At present it has only been possible to fit models for single species and life stages. It is suggested that future work considers interactions between species and life stages and further develops capacity to fit multiple river catchments with the aim of producing a single spatial model for salmonids. Furthermore, there is a need for improved model diagnostics to check model fits and identify influential outliers.

Although the existing datasets were spatially extensive, there is poor replication between Organisation and HA (or other spatial metrics). It is therefore possible that the effects of e.g., equipment or methods are confounded with spatial effects. This could be further investigated if data providers were to fish outside of their geographic areas in locations with similar GIS characteristics. There was also poor data coverage in the case of certain unusual (and broadly correlated) covariate combinations and for larger rivers with a high UCA and river width. Strategic data collection, for example, in areas of high upstream catchment area with low distance to sea (and vice versa); and low channel width with low distance to sea would improve spatial models and help to separate the effects of these variables. Potentially valuable sites could be identified using approaches

similar to those established for the Scottish River Temperature Monitoring Network (SRTMN) (http://www.gov.scot/Topics/marine/Salmon-Trout-Coarse/Freshwater/Monitoring/temperature).

The generation of covariates in ARC GIS was a major resource constraint in the current project. It is also anticipated as a constraint for future model development and predictions where there are potentially even greater data requirements. Increasingly sophisticated spatial tools are being developed in R which allow for rapid automated generation of certain covariates. It is therefore recommended that skills and knowledge are developed in this area.

In the current MasterMap dataset, any waterbodies <1m (urban) or 2m (rural) are characterised using line data. Previous studies (e.g., Wyatt, 2005) have attempted to generate models of channel width from landscape covariates. Although not a priority, similar work could be undertaken for Scotland, with electrofishing site data being used to calibrate these models.

The development of a capture probability model allows for the integration of single and multi-pass electrofishing data on the assumption that they were both collected in a consistent way (i.e., same effort, equipment and staff). Were single pass data to be included in future models, this would greatly increase the size and coverage of available covariates given the quantity of one-pass data collected by fisheries trusts. To ensure that the capture probability models remain valid, it is recommended that multi-pass electrofishing data is still collected by all data providers. Where there is only one-pass data available for particular catchments it is recommended that this is supplemented with multi-pass data from which to estimate capture probability. In addition, given the effect of Organisation on catch probability models, it is likely that further improvements could be made by incorporating information on fishing team and equipment used. It is recommended that the fished area is always recorded as there is no way of incorporating timed fishings.

Given the potential of biomass models to integrate across species and life stages, it would be useful if the sizes of fish (length) are recorded in future electrofishing work in addition to numbers. Finally, reliable estimates of juvenile ages, obtained via scale reading, would allow for the development of more

complex age structured models which potentially indicate additional system resilience over and above numbers alone.

A substantial amount of resource was required to collate and harmonise data formats for the electrofishing data and also in allocating recorded sites to locations on rivers. Future projects would benefit greatly if these data were collected and stored using the same data format, preferably in a single database. MSS has found that the use of a single flexible database reduces the need for multiple extractions, followed by complex collation, and can allow for validation of site locations, removing some of the potential errors associated with snapping sites to rivers.

Finally, the selection of covariates in the current report was pragmatic given the need for large scale spatial coverage and the limited time available for the project. However, future work should consider the inclusion of metrics of hydrological impact, hydrochemistry and river temperature given the importance of these variables in controlling fish distribution and abundance. Due to the discrete nature of water quality observations, these data would need to be modelled using similar approaches to those reported here.

Acknowledgements

The authors thank the SFCC and associated fishery trusts in the following areas that provided data used in this report: Annan, Argyll, Ayr, Clyde, Conon, Deveron, Esk, Findhorn, Forth, Galloway, Kyle of Sutherland, Lochaber, Ness and Beauly, Outer Hebrides, Dee, Spey, Tay, Tweed, West Sutherland, Wester Ross, Naver and Nith. The data contributions of SEPA, Alan Youngson and MSS staff are also gratefully acknowledged. We thank Sean Dugan for liaising with trust members and exporting data from the SFCC database.

References

Armstrong, J.D., Kemp, P.S., Kennedy, G.J.A., Ladle, M. and Milner, N.J. 2003. Habitat Requirements of Atlantic Salmon and Brown Trout in Rivers and Streams. *Fisheries Research*, **62** (2): 143–70.

Bohlin, T., Pettersson J. and Degerman E. 2001. Population Density of Migratory and Resident Brown Trout (*Salmo Trutta*) in Relation to Altitude: Evidence for a Migration Cost. *Journal of Animal Ecology*, **70** (1): 112–21.

Borgstrøm, R. and Skaala Ø. 1993. Size-Dependent Catchability of Brown Trout and Atlantic Salmon Parr by Electrofishing in a Low Conductivity Stream. *Nordic Journal of Freshwater Research*, **68**: 14–21.

Cressie, N., Frey, J., Harch, B. and Smith, M. 2006. Spatial Prediction on a River Network. *Journal of Agricultural, Biological, and Environmental Statistics*, **11** (2): 127–50.

Deschênes, J. and Rodríguez M.A. 2007. Hierarchical Analysis of Relationships between Brook Trout (*Salvelinus Fontinalis*) Density and Stream Habitat Features. *Canadian Journal of Fisheries and Aquatic Sciences*, **64** (5): 777–85.

Fahrmeir, L., Kneib, T., Lang, S. and Marx, B. 2013. Regression. Berlin, Springer.

Fausch, K.D., Hawkes, C.L. and Parsons, M.G. 1988. Models That Predict Standing Crop of Stream Fish from Habitat Variables: 1950-85. *Gen. Tech. Rep. PNW-GTR-213*. U.S. Department of Agriculture.

Godfrey, J.D. 2005. Site Condition Monitoring of Atlantic Salmon SACs. Report by the SFCC to Scottish Natural Heritage, Contract F02AC608, 274 pp.

Kennedy, G.J.A. and Strange, C.D. 1981. Efficiency of Electric Fishing for Salmonids in Relation to River Width. Aquaculture Research, **12**: 55–60.

Lanka, R.P., Hubert, W.A. and Wesche, T.A. 1987. Relations of Geomorphology to Stream Habitat and Trout Standing Stock in Small Rocky Mountain Streams. *Transactions of the American Fisheries Society*, **116** (1): 21–28.

Marsh, T. J. and Hannaford, J. (Eds). 2008. UK Hydrometric Register. Hydrological data UK series. Centre for Ecology & Hydrology. 210 pp.

Moran, P.A.P. 1951. A Mathematical Theory of Animal Trapping. *Biometrika*, **38**: 307–11.

Niemelä, E., Julkunen, M. and Erkinaro, J. 2000. Quantitative Electrofishing for Juvenile Salmon Densities: Assessment of the Catchability during a Long-Term Monitoring Programme. *Fisheries Research*, **48** (1): 15–22.

Otis, D.L., Burnham, K.P., White, G.C. and Anderson, D.R. 1978. Statistical Inference from Capture Data on Closed Animal Populations. *Wildlife Monographs*, **62**: 3–135.

Peterson, E.E., Ver Hoef, J.M., Isaak, D.J., Falke, J.A., Fortin, M.J., Jordan, C.E., McNyset, K., *et al.* 2013. Modelling Dendritic Ecological Networks in Space: An Integrated Network Perspective. *Ecology Letters*, **16** (5): 707–19.

Prévost, E., Parent, E., Crozier, W., Davidson, I., Dumas, J., Gudbergsson, G., Hindar, K., McGinnity, P., MacLean, J. and Sættem, L.M. 2003. Setting Biological Reference Points for Atlantic Salmon Stocks: Transfer of Information from Data-Rich to Sparse-Data Situations by Bayesian Hierarchical Modelling. *ICES Journal of Marine Science: Journal Du Conseil*, **60** (6): 1177–93.

Rivot, E., Prévost, E., Cuzol, A., Baglinière, J.-L. and Parent, E. 2008. Hierarchical Bayesian Modelling with Habitat and Time Covariates for Estimating Riverine Fish Population Size by Successive Removal Method. *Canadian Journal of Fisheries and Aquatic Sciences*, **65** (1): 117–33.

Rosenfeld, J., Porter, M. and Parkinson, E. 2000. Habitat Factors Affecting the Abundance and Distribution of Juvenile Cutthroat Trout (*Oncorhynchus Clarki*) and Coho Salmon (*Oncorhynchus Kisutch*). *Canadian Journal of Fisheries and Aquatic Sciences*, **57** (4): 766–74.

Rue, H. and Held, L. 2005. Gaussian Markov Random Fields. Monographs on Statistics and Applied Probability, 104. Boca Raton, Chapman and Hall / CRC press.

Schwarz, G. 1978. Estimating the dimension of a model. *The annals of statistics*, **6** (2): 461-464.

SNIFFER. 2011. River Fish Classification Tool: Science Work, Phase 3 Report, Final. Project WFD68c. Scottish and Northern Ireland Forum for Environmental Research.

Wood, S.N. 2006. Generalized additive models: an introduction with R. CRC press.

Wood, S.N. 2011. Fast Stable Restricted Maximum Likelihood and Marginal Likelihood Estimation of Semiparametric Generalized Linear Models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **73** (1): 3–36.

Wyatt, R.J. 2002. Estimating Riverine Fish Population Size from Single- and Multiple-Pass Removal Sampling Using a Hierarchical Model. *Canadian Journal of Fisheries and Aquatic Sciences*, **59** (4): 695–706.

Wyatt, R.J. 2003. Mapping the Abundance of Riverine Fish Populations: Integrating Hierarchical Bayesian Models with a Geographic Information System (GIS). *Canadian Journal of Fisheries and Aquatic Sciences*, **60** (8): 997–1006.

Wyatt, R.J. 2005. River Fish Habitat Inventory Phase 2 : Methodology Development for Juvenile Salmonids. Science Report SC980006. Environment Agency.

Wyatt, R.J. and Barnard, S. 1997. The transportation of the maximum gain salmon spawning target from the River Bush (N.I.) to England and Wales. *R & D Technical Report W65*. Environment Agency.

Wyatt, R.J., Sedgwick, R. and Burrough, R. 2007a. A Statistical Approach to the Assessment of Coarse Fish Populations. Science Report SC030214. Environment Agency.

Wyatt, R.J., Sedgwick, R. and Simcox, H. 2007b. River Fish Habitat Inventory Phase 3: Multi-Species Models. Science Report SC040028. Environment Agency.

Zippin, C. 1956. An Evaluation of the Removal Method of Estimating Animal Populations. *Biometrics*, **12** (2): 163–89.

Marine Scotland Science
Freshwater Laboratory
Faskally
Pitlochry
PH16 5LB

Copies of this report are available from the Marine Scotland website at http://www.gov.scot/marinescotland

This publication is available at **www.scotland.gov.uk**

Any enquiries regarding this publication should be sent to us at
The Scottish Government
St Andrew's House
Edinburgh
EH1 3DG

ISBN: 978-1-78544-384-8 (web only)

Published by The Scottish Government, May 2015

Produced for The Scottish Government by APS Group Scotland, 21 Tennant Street, Edinburgh EH6 5NA
PPDAS50381 (05/15)