Mapping habitat indices across river networks using spatial statistical modelling of River Habitat Survey data.

Marc Naura[1], Mike J. Clark[2] (†), David A. Sear[2], Peter M. Atkinson[3], Duncan D. Hornby[2], Paul Kemp[1], Judy England[4], Graeme Peirson[5], Chris Bromley[6], Matthew G. Carter[7]

[1] Faculty of Engineering and the Environment, University of Southampton, SO17 1BJ, Southampton, UK; [2] School of Geography, University of Southampton, Southampton, SO17 1BJ, UK; [3] Faculty of Science and Technology, Engineering Building, Lancaster University, Bailrigg, Lancaster LA1 4YR, UK; [4] Environment Agency, Red Kite House, Howbery Park, Crowmarsh Gifford, Wallingford OX10 8BD, UK; [5] Mance House, Arthur Drive, Hoo Farm Industrial Estate, Worcester Road, Kidderminster DY11 7RA, UK. [6] Scottish Environment Protection Agency, Strathallan House, Castle Business Park, Stirling FK9 4TZ, UK; [7] Environment Agency, Thames East, Apollo Court, 2 Bishops Square, St. Albans, Road West, Hatfield, Hertfordshire AL10 9EX, UK.

Corresponding author: Marc Naura, marc.naura@soton.ac.uk, Faculty of Engineering and the Environment, University of Southampton, SO17 1BJ, Southampton; tel: + 44 2380 592700

Abstract

Freshwater ecosystems are declining faster than their terrestrial and marine counterparts because of physical pressures on habitats. European legislation requires member states to achieve ecological targets through the effective management of freshwater habitats. Maps of habitats across river networks would help diagnose environmental problems and plan for the delivery of improvement work. Existing habitat mapping methods are generally time consuming, require experts and are expensive to implement. Surveys based on sampling are cheaper but provide patchy representations of habitat distribution. In this study, we present a method for mapping habitat indices across networks using semi-quantitative data and a geostatistical technique called regression kriging. The method consists of the derivation of habitat indices using multivariate statistical techniques that are regressed on map-based covariates such as altitude, slope and geology. Regression kriging combines the Generalised Least Squares (GLS) regression technique with a spatial analysis of model residuals. Predictions from the GLS model are 'corrected' using weighted averages of model residuals following an analysis of spatial correlation. The method was applied to channel substrate data from the River Habitat Survey in Great Britain. A Channel Substrate Index (CSI) was derived using Correspondence Analysis and predicted using regression kriging. The model explained 74% of the main sample variability and 64% in a test sample. The model was applied to the English and Welsh river network and a map of CSI was produced. The proposed approach demonstrates how existing national monitoring data and geostatistical techniques can be used to produce continuous maps of habitat indices at the national scale.

Keywords: habitat mapping, habitat indices, channel substrate, regression kriging, River Habitat Survey, geostatistics

1. Introduction

Freshwater ecosystems represent less than 1% of the Earth's surface and 10% of all known species, yet they are declining faster and are more endangered than their terrestrial or marine counterparts, partly because of physical pressures on habitats and species (Loh et al., 2005; Revenga et al., 2005; Strayer and Dudgeon, 2010; Vorosmarty et al., 2010; WWF, 2014).

Although research in ecology and environmental management has grown substantially in the past half-century, it has mainly focused on post-industrial issues such as water quality, pollution and land use impacts (Vaughan et al., 2009). With gradual improvement in water quality, other limiting factors such as physical habitat quality (i.e. the naturalness of the flow of water, and the structure and composition of the river bed and banks) and connectivity have become prominent.

Globally, degradation of physical habitat quality due to river engineering and associated activities (e.g. constructions of dams, bridges, concrete banks, dredging) is recognised as a major conservation issue (Collen et al., 2014; Sala et al., 2000; Tockner and Stanford, 2002; World Conservation Monitoring Centre, 1998). In Europe, as part of the implementation of the Water Framework Directive (WFD), member states must assess the ecological condition of rivers and lakes based on the naturalness of a series of biological elements (European Union, 2000). Following the first round of River Basin Management Planning, 56% of water bodies failed to achieve their ecological targets. Engineered structures and 'altered habitats' were the dominant pressures responsible for the failure, ahead of point and diffuse sources of pollution (European Environment Agency, 2012). In England and Scotland, the proportion of water bodies failing to achieve ecological targets because of physical alterations was 49% and 37%, respectively (Environment Agency, 2012). The WFD requires member states to mitigate or remove impacts on habitats and species through the implementation of programmes of measures including river restoration.

The effective management of habitats at global and local scales should ideally be based on some knowledge of their distribution and an assessment of their naturalness and accessibility. At present, in Great Britain, habitats are either surveyed using semi-quantitative methods at randomly selected sites that do not allow for continuous assessments or using habitat mapping techniques over longer stretches of river (Maddock, 1999). Habitat mapping is geographically limited and generally carried out on an *ad hoc* basis by experts during 'walkover surveys' where habitat features are recorded on maps using mobile Geographic Information System (GIS) or hand-drawn sketches and some broad typologies (Hendry and Cragg-Hine, 1997; Sear et al., 2009). Although such methods provide valuable information on habitat distributions over relatively small areas, they are likely to be too expensive to implement across entire networks. The reliance on expert judgement for assessing habitat types and boundaries may also generate between-surveyor variability in the outputs produced and, as notions of habitat structure evolve, data collected at one point in time may not be comparable to maps produced years later by different experts (Cherrill and Mcclean, 1999).

An alternative approach is to use river typologies based on geomorphological templates to predict the occurrence of broad river types along the river continuum. The history of attempts to classify rivers into different types spans at least 125 years, a period over which perhaps a hundred if not more individual efforts to divide and categorise rivers have been made (reviews of the extent of such efforts are given by Downs, 1995; Montgomery and Buffington, 1997; Mosley, 1987; Naiman et al., 1992; Newson et al., 1998; Thorne, 1997).

Most river classification systems are based on the identification of river types using a few key variables representing drivers of geomorphological change or river processes such as stream power, sediment transport and supply (Montgomery and Buffington, 1997; Newson et al., 1998; Rosgen, 1994). Although relationships between expert-driven geomorphic types and GIS attributes such as slope and drainage area can be observed, there is a considerable amount of overlap between types, reflecting the potential influence of additional driving elements such as channel, bank and hillslope vegetation, climate, woody debris, and natural variability in channel process expression (Church, 2002; Montgomery and Buffington, 1997; Rosgen, 1994). Greater differentiation between river types can be achieved by introducing attributes recorded in the field such as relative roughness (Montgomery and Buffington, 1997), shear stress or channel substrate (Rosgen, 1994), but this implies that extensive field work is carried out, thus reducing the feasibility of such an approach at national scales.

In this article, we propose an alternative approach for mapping habitat elements across entire river networks that does not require continuous surveys of river catchments, but makes use of existing semi-quantitative survey data, GIS and a geostatistical technique called regression kriging (RK). The principle of the method is to identify and define habitat indices representing major dimensions in habitat distribution using known equations, expert systems or multivariate statistical analysis applied to existing habitat data taken from national surveys or monitoring programmes. The habitat indices are then predicted using Generalised Least Squares (GLS) linear regression models using GIS map-derived covariates such as altitude, slope, distance from source, discharge and geology which represent the known drivers of habitat/geomorphological change. The model residuals are then analysed using geostatistical functions to identify any remaining spatial correlation and pattern in their distribution. In the presence of spatial correlation, an interpolation method, called kriging, is applied to account for (and, thus remove) any spatially correlated residual variance such that the interpolated residual predictions can be added to the GLS regression predictions. The RK model can then be applied to the entire river network by deriving the GIS covariates at regular spatial intervals (e.g. 500 m).

This paper reports the development and application of the statistical models to a key and poorly mapped habitat element – channel substrate. Channel substrate is a key component of species habitat (Maddock, 1999; Townsend and Hildrew, 1994), and it is one of three elements defining morphological condition under the WFD (European Union, 2000). Channel substrate is also linked to the wider issues of diffuse pollution and agricultural impacts and it is key to our understanding of river and catchment processes (Collins et al., 2014; Rosgen, 1994).

2. Material and methods

2.1. Index derivation

River Habitat Survey (RHS) data was used to derive an index representing channel substrate. RHS is a CEN-compliant (CEN, 2004) standard methodology for hydromorphological assessment under the WFD that is used in the UK and across Europe (Raven et al., 1997). It is a methodology for recording habitat features for wildlife that has been implemented at more than 25,000 sites in the UK since 1994. From 1994 to 1996 and from 2007 to 2008, surveys were carried out at random sites in every 10 km^2 in England and Wales, thus, ensuring a wide geographical coverage of the river network.

RHS records the presence of natural and management features at 10 equally spaced transects or 'spot-checks' along a 500 m reach (Raven et al., 1997). A visual estimate of the dominant channel surface substrate classified into eight categories according to the Wentworth scale (Wentworth, 1922) is recorded at each spot-check. The substrate types recorded (with acronyms in brackets) are bedrock (BE), boulder (BO), cobble (CO), gravel-pebble (GP), sand (SA), silt (SI), clay (CL) and peat (PE). When channel substrate is not visible because of depth, water turbidity or the presence of a culvert, surveyors record the substrate type as ‘Not Visible’ (NV).

RHS spot-check data on channel substrate was tabulated for all existing sites, each row representing a site and each column a substrate type (including ‘Not visible’). The channel substrate spot-check table was analysed using Correspondence Analysis (CA). CA is a multivariate analytical technique similar to Principal Component Analysis that is applicable to contingency tables (i.e. tables of counts). CA performs an analysis of the total table inertia and extracts dimensions (or components) representing linear combinations of input variables based on the amount of total inertia explained. Only sites in Great Britain were used as GIS datasets were not available for Northern Ireland at the time of the analyses.

To derive the index, we used a subset of 2680 semi-natural RHS sites (i.e. sites with few or no in-channel bank structures or modifications) to reduce the potential influence of modifications on natural channel substrate diversity (Raven et al., 1997). Missing (‘Not Visible’) values were added as an additional variable in the analyses to account for differences in survey counts when present. The resulting dimensions were investigated for their ecological and geomorphological significance and for the amount of variability (i.e. inertia) they explained. One dimension was chosen to represent substrate and calculate an index score, called the Channel Substrate Index (CSI) for all sites in the RHS database.

2.2. Regression kriging

RK was applied following an iterative procedure using both Ordinary Least Square (OLS) and GLS regression techniques (Bivand et al., 2008; Webster and Oliver, 2007). The CSI index was first transformed using a Box-Cox procedure and modelled against a series of GIS attributes: four Principal Component Axes (PCA) combining altitude, slope, distance to source and height of source that were shown by Jeffers (1998) and Vaughan et al. (2013) to be strongly correlated to sediment distribution; land use categories from the Land Cover Map 2000 (Fuller et al., 2002); British Geological Survey solid and drift geology categories taken from the 1/625,000 scale maps; hydrometric areas corresponding to large catchment areas; and solid geology age categorised in 11 groups from the pre-Cambrian to the Neogene. Solid geology age was included as a surrogate for hardness as older rocks tend to be harder and display coarser substrate types than softer and younger sedimentary deposits.

Nominal attributes such as solid geology, hydrometric area and land use were transformed into binary indicator variables. Due to the resulting large number of indicator variables which would have rendered the predictive models difficult to display and interpret (e.g. there are more than 100 different solid geology types), indicator variables were grouped based on their relationships to the CSI. Grouping was done by comparing coefficient values of indicator variables when individually regressed against CSI or performing ANOVAs.

Only RHS sites with no missing channel substrate spot-check records were retained for the analysis as their presence introduces a potential bias in channel substrate representation and prediction. Model selection was performed using the Minitab 16 (Minitab, 2010) linear regression (OLS) best subset selection procedure using Mallows Cp (all models) on RHS sites from 1994 to 2005 (9473 sites).

Model residuals were analysed for the presence of spatial correlation using a variogram (Webster and Oliver, 2007), which plots semivariance (a measure of dissimilarity) against lag vector (the distance and direction of separation). Spatially uncorrelated data display no observable change in semivariance with an increase in lag distance and are typically represented by a flat variogram. Spatially correlated data are typically represented by a monotonically increasing semivariance as the lag distance between sites increases.

The empirical variogram is first calculated as the average squared difference between pairs of data points at each of a series of lags. It is fitted with a permissible variogram model (the model must not result in negative prediction variances) to describe the shape of the curve and identify the parameters which are required for RK. Of particular relevance are the nugget and the range parameters (assuming that a bounded model such as the spherical or exponential model is fitted). The nugget variance is equal to the variance for sites re-surveyed or re-sampled at the same location and expresses micro-scale variability and survey error. The range is the distance at which the semivariance reaches a plateau and beyond which data are no longer spatially correlated (Webster and Oliver, 2007).

The OLS variogram parameters were used as part of an iterative process to estimate the regression parameters using GLS. GLS is preferred to OLS as the latter assumes independence of observations, an unlikely case given spatial correlation (Bivand et al., 2008). Model residuals were checked for the presence of trends and outliers. The models were validated using a leave-one-out cross-validation technique (Bivand et al., 2008; Vaughan and Ormerod, 2003).

The model was applied to all points using the GSTAT kriging procedure. Kriging linearly averages the residual values of surrounding points with weights estimated using the residual variogram (Webster and Oliver, 2007). To reduce processing time, only points within a radius roughly equal to the distance at which no observable correlation exists (i.e. the range) were selected for kriging. The kriged residuals were then saved and added to the cross-validated predictions from the previous model. A pseudo-R^2 value for the final predictions was derived by correlating the predicted and observed values. Residuals were computed and checked for their distribution and for signs of remaining spatial correlation. A check of model stability against time was performed by examining the residual distribution for each year of survey. The model was tested on a sample of 3884 sites collected from 2006-2011.

The predictive model and kriging procedure were applied to the entire English and Welsh 1:50,000 river network to produce a national map of sediment distribution. To do so, points were generated every 500 m on the river network using RivEX (Hornby, 2010) and the GIS map-based covariates required for prediction were derived for each point.

3. Results

3.1. Channel Substrate Index

The first two components of the CA explained 21% and 17% of the total inertia (Table 1). The first component represented a gradient between sites dominated by fine substrate such as silt, clay and sand, and sites dominated by coarse substrate such as bedrock and boulders (Fig. 1). The first component was defined by the relative occurrence of all substrate types with a greater contribution from silt which explained 35% of the component inertia (Table 1). The first component explained 42% of silt distribution inertia, 31% of boulder inertia, 23% for cobbles and 20% for gravel pebble and bedrock.

Table 1: Simple CA on channel substrate types for 2680 semi-natural RHS in Great Britain. Only detailed results for the first 2 components are displayed. The 'Coord' columns contain the principal coordinates for each substrate type and axis. The 'Contr' column expresses the relative contribution of individual substrate types to axis definition whilst the 'Corr' column (or relative contribution) represents the amount of individual substrate inertia explained by each component (Greenacre, 1993).

Individual axes inertia relative to total inertia

Axis	Inertia	Proportion	Cumulative
1	0.7767	0.2051	0.2051
2	0.6442	0.1701	0.3752
3	0.5991	0.1582	0.5334
4	0.5632	0.1487	0.6821
5	0.4317	0.1140	0.7961
6	0.4134	0.1092	0.9052
7	0.3589	0.0948	1.0000
Total	3.7872		

Column Contributions for components 1 and 2

	Component .1			Component 2		
Name	Coord	Corr	Contr	Coord	Corr	Contr
BE	0.793	0.190	0.097	-0.229	0.016	0.010
BO	0.835	0.307	0.151	-0.230	0.023	0.014
CO	0.511	0.226	0.097	-0.077	0.005	0.003
GP	-0.532	0.197	0.106	0.394	0.108	0.070
SA	-1.433	0.160	0.125	2.480	0.480	0.449
SI	-2.052	0.423	0.349	-2.119	0.451	0.449
CL	-2.008	0.097	0.074	-0.483	0.006	0.005
PE	0.067	0.000	0.000	0.039	0.000	0.000

Supplementary Columns

	Component 1			Component 2		
Name	Coord	Corr	Contr	Coord	Corr	Contr
NV	-0.540	0.014	0.044	-0.084	0.000	0.001

The first CA axis represented a gradual increase in substrate size with a gradual shift from sites dominated by fine sediment to sites dominated by larger substrate (Fig. 2). The second component represented a gradient between silt and sand dominated sites and explained nearly 50% of the inertias of both substrate types (Table 1). The remaining components either represented gradients between two or three substrate types or were linked to the occurrence of rare types such as peat or clay. Missing values were not associated with any particular substrate category and only 1% of the missing values inertia was explained by the first two components.

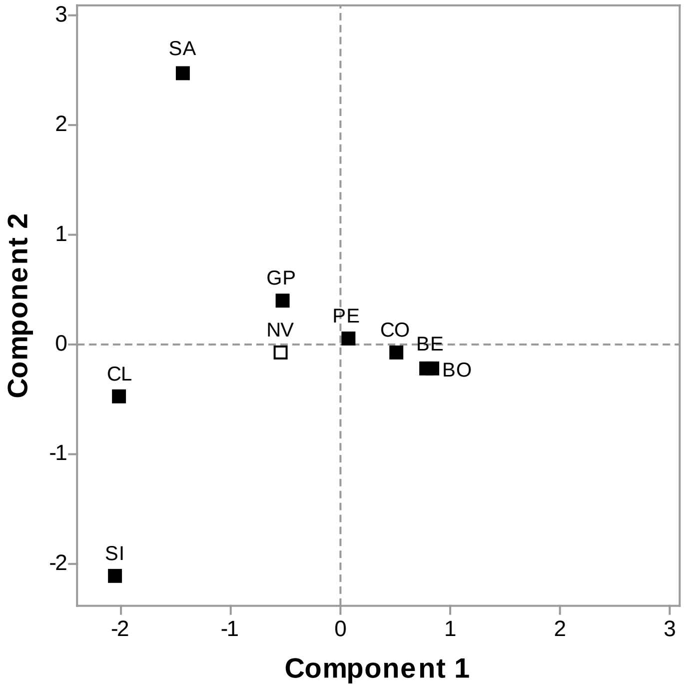

Fig. 1: Symmetrical plot of substrate category profiles for the first two CA axes.

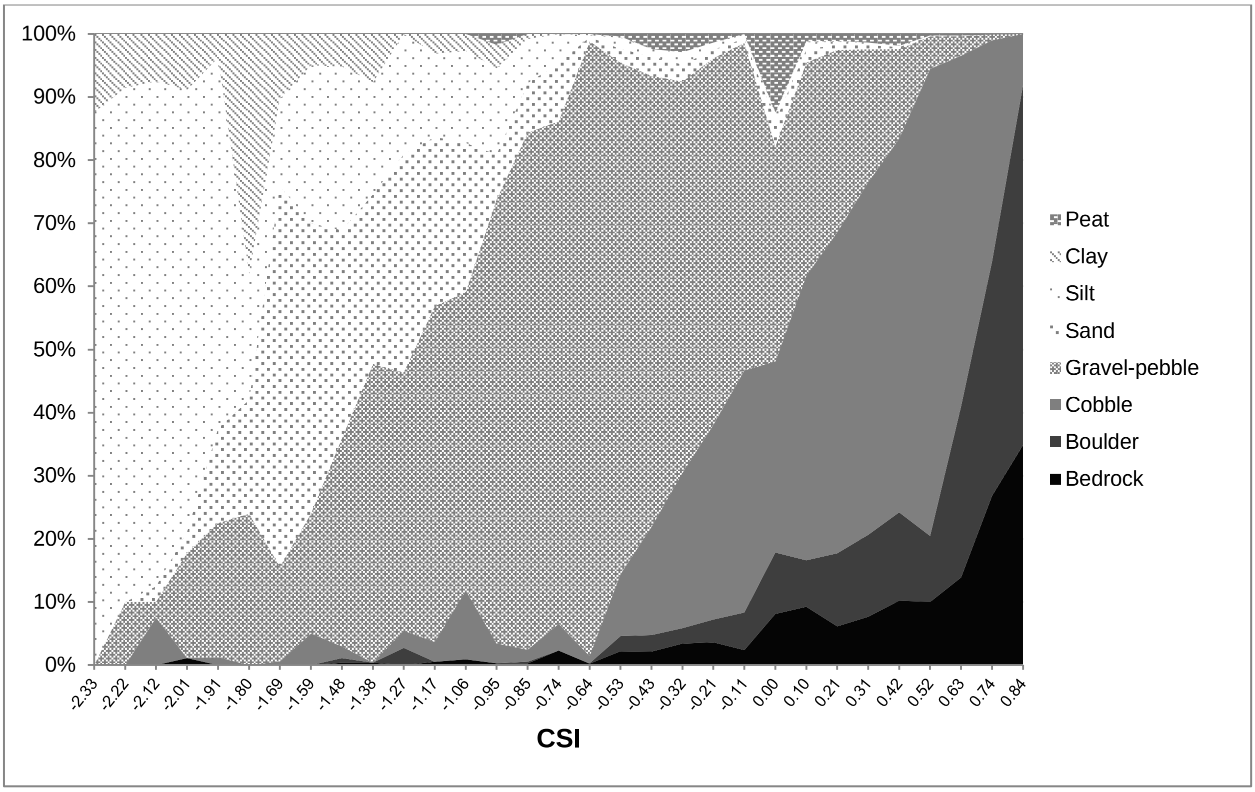

Fig. 2: RHS sites were grouped into 31 bins based on their CSI index values. The plot displays for each bin the average occurrence of channel substrate types.

The first component was chosen for its geomorphological relevance as it represented a well-known dimension in sediment fining and sorting along the river network (Morris and Williams, 1999) and has habitat significance with regards to species distribution (Chessman et al., 2006; Gasparini et al., 1999; Rice et al., 2001). The CSI was calculated for all existing RHS sites using channel substrate standard coordinates for the first component in the following equation:

$$CSI = (0.89(AR+BE) + 0.95\ BO + 0.58\ CO + 0.08\ PE - 0.6\ GP - 1.63\ SA – 2.33\ SI – 2.28\ CL) / N_{sc}$$

where each two-letter acronym refers to RHS channel substrate categories and N_{sc} is the total number of spot-checks. Artificial channel substrate (AR) was given the same coefficient as bedrock substrate.

3.2. Variable selection

Only attributes selected by the best subset procedure will be presented and discussed. Land use categories and drift geologies were not selected in any of the models extracted using the best subset procedure. The attributes retained for the analyses were the PCA axes, solid geology age and categories and hydrometric areas.

The four PCA variables represent environmental gradients describing site location and profile (PCA1; i.e. lowland low altitude and slope, and upland high altitude and slope), catchment area (PCA2), local discontinuities in profile/geology (PCA3) and catchment slope (PCA4) (Jeffers, 1998).

Solid geologies were split into two groups based on their age. More recent geologies from the Permian and Triassic to the Neogene had significantly lower CSI values indicating finer substrate types than geologies from the Carboniferous to the Precambrian (Fig. 3). The geology age categories were recoded into one indicator variable coding for solid geology ages younger than the Carboniferous (Fig. 4). Geology age distribution separates Wales, Cornwall and part of the North and the Lake District from the lowland areas of eastern and southern England.

Hydrometric areas were recombined into six groups based on their average CSI value and ordered according to increasing substrate size. The distribution of hydrometric area categories follows a pattern separating the uplands from the lowlands of England and Wales (Fig. 5). Hydrometric area categories represent catchment size and its influence on substrate with smaller hydrological units displaying coarser substrate types than larger lowland catchments. Group 4, 5 and 6 represent steeper catchments with higher levels of hill slope activity and reflect the preponderance of upland controls in the delivery of sediments. Catchments from the lower groups tend to have a higher proportion of streams originating and running in low altitude low slope areas compared to higher categories. Coarse sediment tends to originate closer to source and is linked to local erosion of hard rocks generally located in the upland areas. A lack of upland control within catchments is, therefore, likely to result in lower delivery of coarse sediments within the river system and a higher proportion of fine sediment arising from downstream attrition and fining (Werritty, 1992).

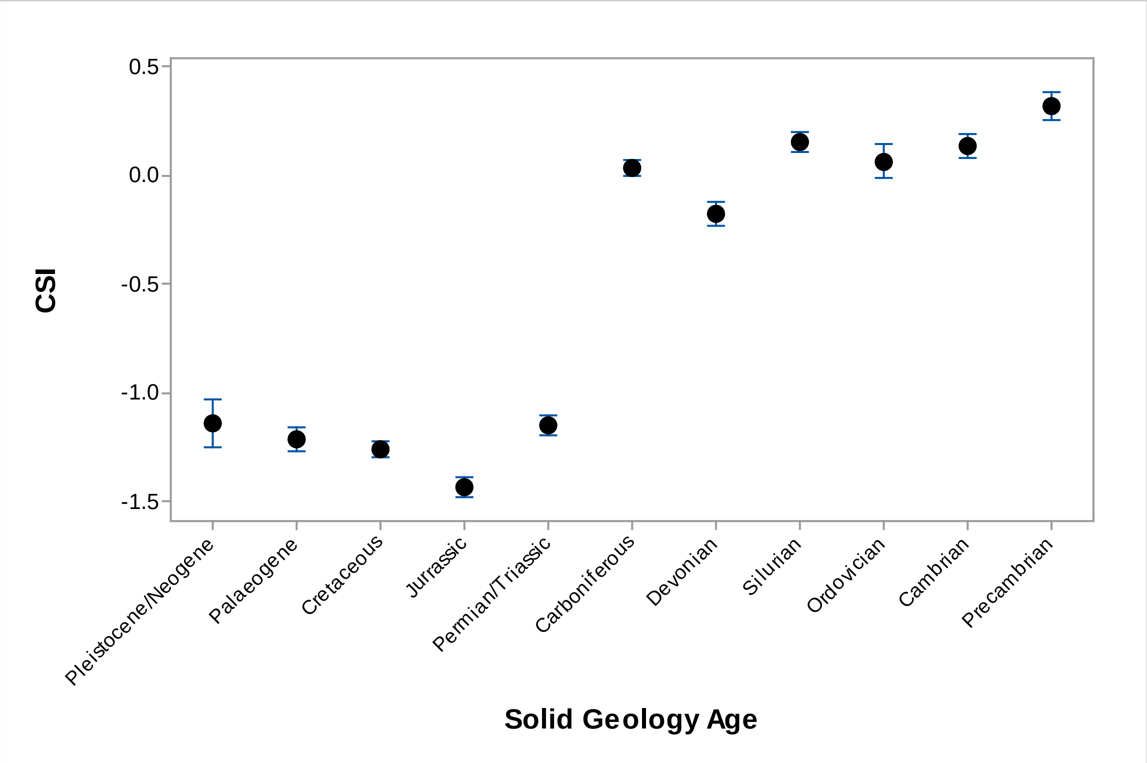

Fig. 3: One way Anova of CSI value against solid geology age categories derived from the 1979 BGS solid geological map for all RHS sites. Average CSI values per age category with 95% confidence intervals based on pooled standard deviation (F=819, p<0.0001, n=9934).

Solid geologies were grouped into eight classes based on increasing average CSI value (Fig. 6). Upon examination, solid geology categories reflect two related aspects: geological age and hardness. The first class contains recent erodible clay and limestone formations and displays rivers with predominantly fine sediment material. The following two solid geology classes comprise slightly older (Jurassic) soft sedimentary formations including chalk, clay, limestone, shales and marls that support streams dominated by fine sediments with some occurrence of gravels and pebbles. Class three is dominated by sedimentary sand, clay and Oolitic geologies. River channels running on those geologies tend to display a higher fraction of gravels and pebbles with a lower predominance of fine sediment. Solid geology class four is constituted of older and harder geologies from the Carboniferous/Triassic period with sandstone and coal that support rivers with a significantly higher occurrence of coarse substrate such as gravel-pebbles and cobbles and little fine sediment. Class five contains geologies from the Cambrian up to the Carboniferous with a predominance of metamorphic and intrusive rocks such as grit stone and granite. Rivers running on these profiles tend to have coarser substrate with cobbles, boulders and bedrock. Class six is constituted mainly of hard igneous and Palaeozoic sedimentary rocks. These are associated with rivers showing a dominance of cobbles with greater occurrence of boulders and bedrock. The last class represents Cambrian grit and limestone rocks that are characterised by very coarse substrate types. Geographically, harder and older geologies are located in the west of England, in Wales, in the North West and near the Scottish border (Fig. 6).

From the previous three maps, we can observe spatial correlations between geological age, solid geology classes and hydrometric area categories. Although the three sets of variables are correlated, they each provide subtle differences in explaining substrate distribution.

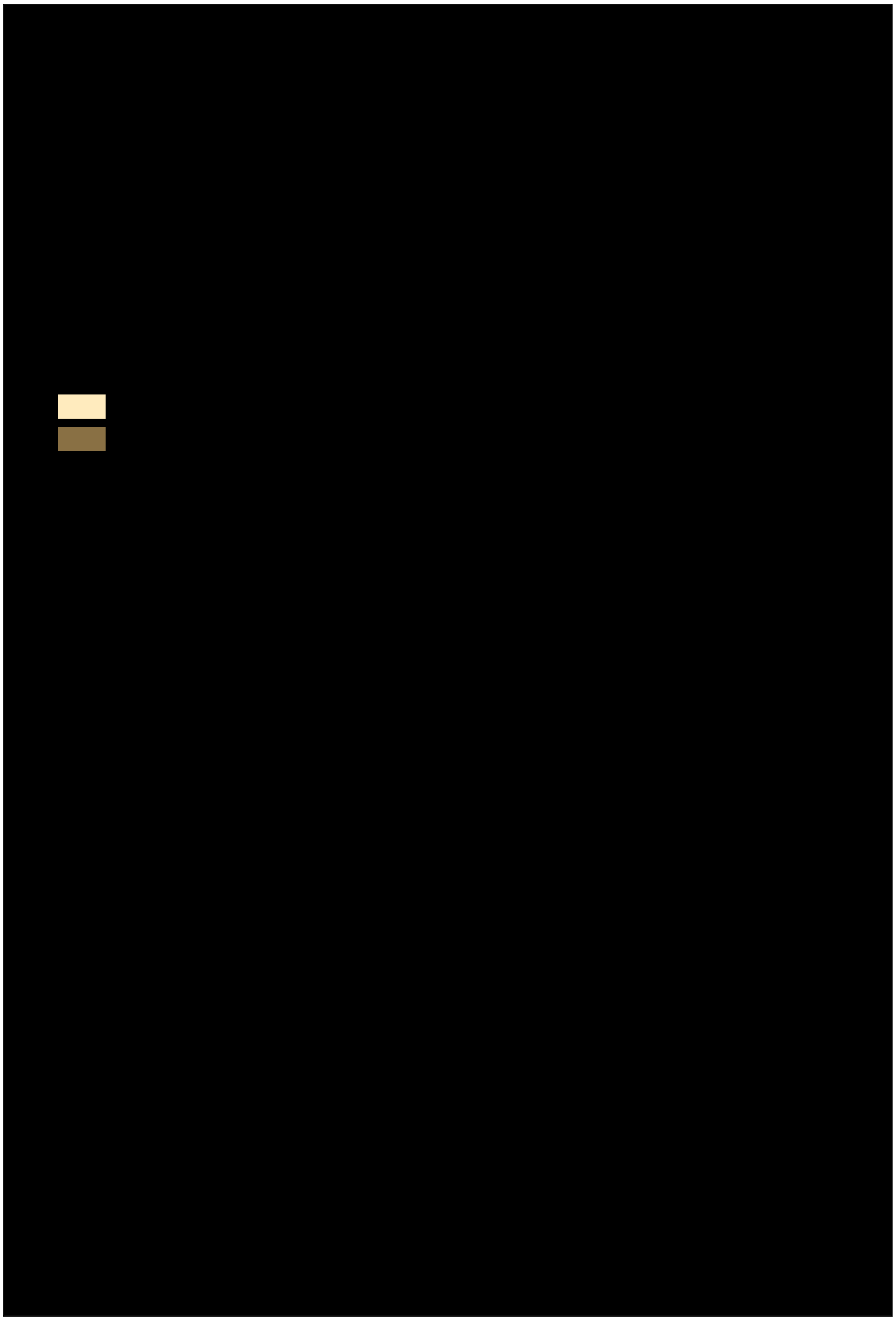

Fig. 4: Solid geology age distribution in England and Wales from the 1:625,000 BGS solid geology map recombined into two classes. ‘Recent geologies’ represent geologies from the Neogene to the Triassic and ‘older geologies’ from the Carboniferous to the Precambrian.

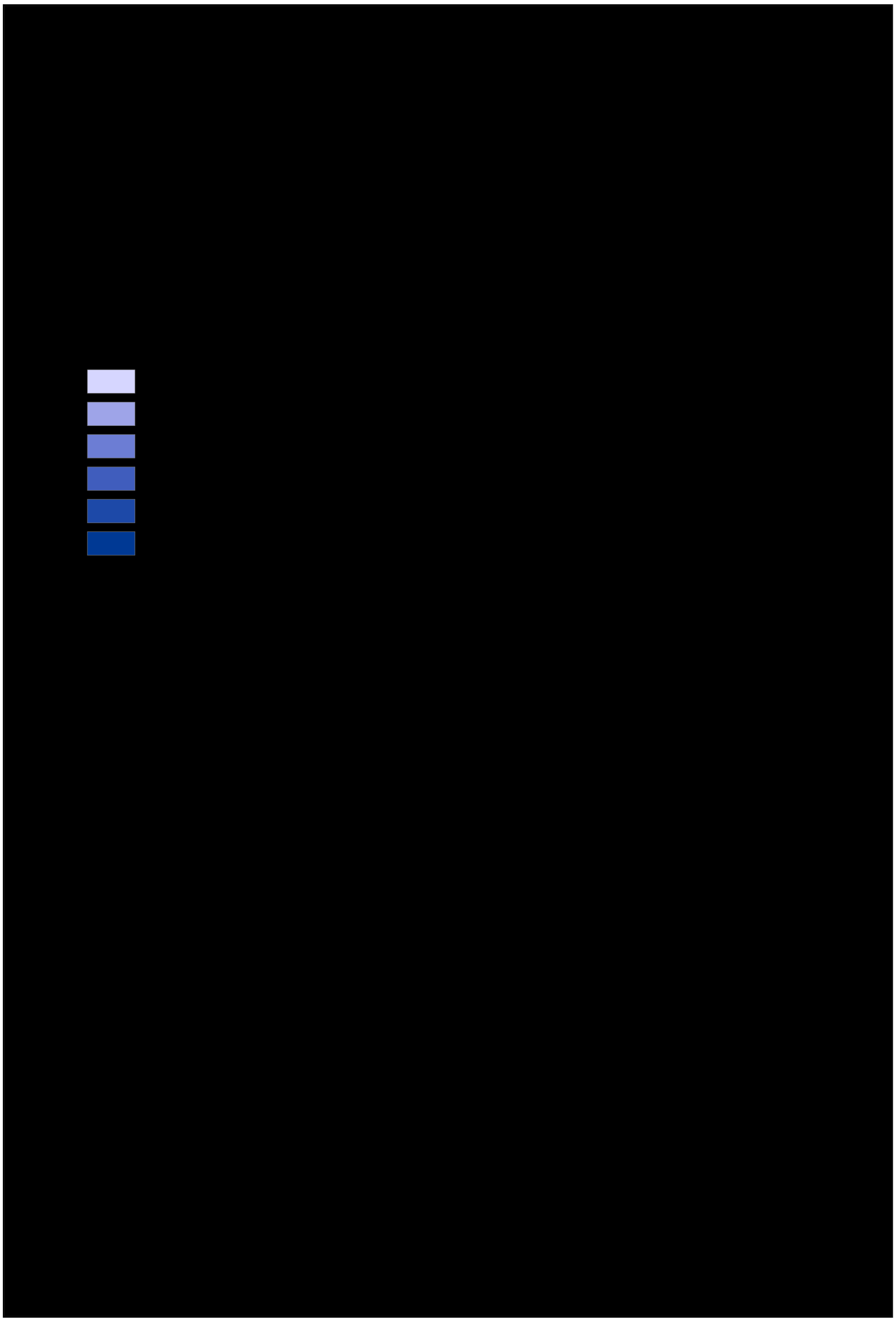

Fig. 5: Hydrometric area category distribution in England and Wales. Categories from 1 to 6 represent hydrological units with increasing average CSI values for surveyed RHS sites. Low CSI values correspond to fine sediment dominated streams, high CSI to coarse sediment dominated streams.

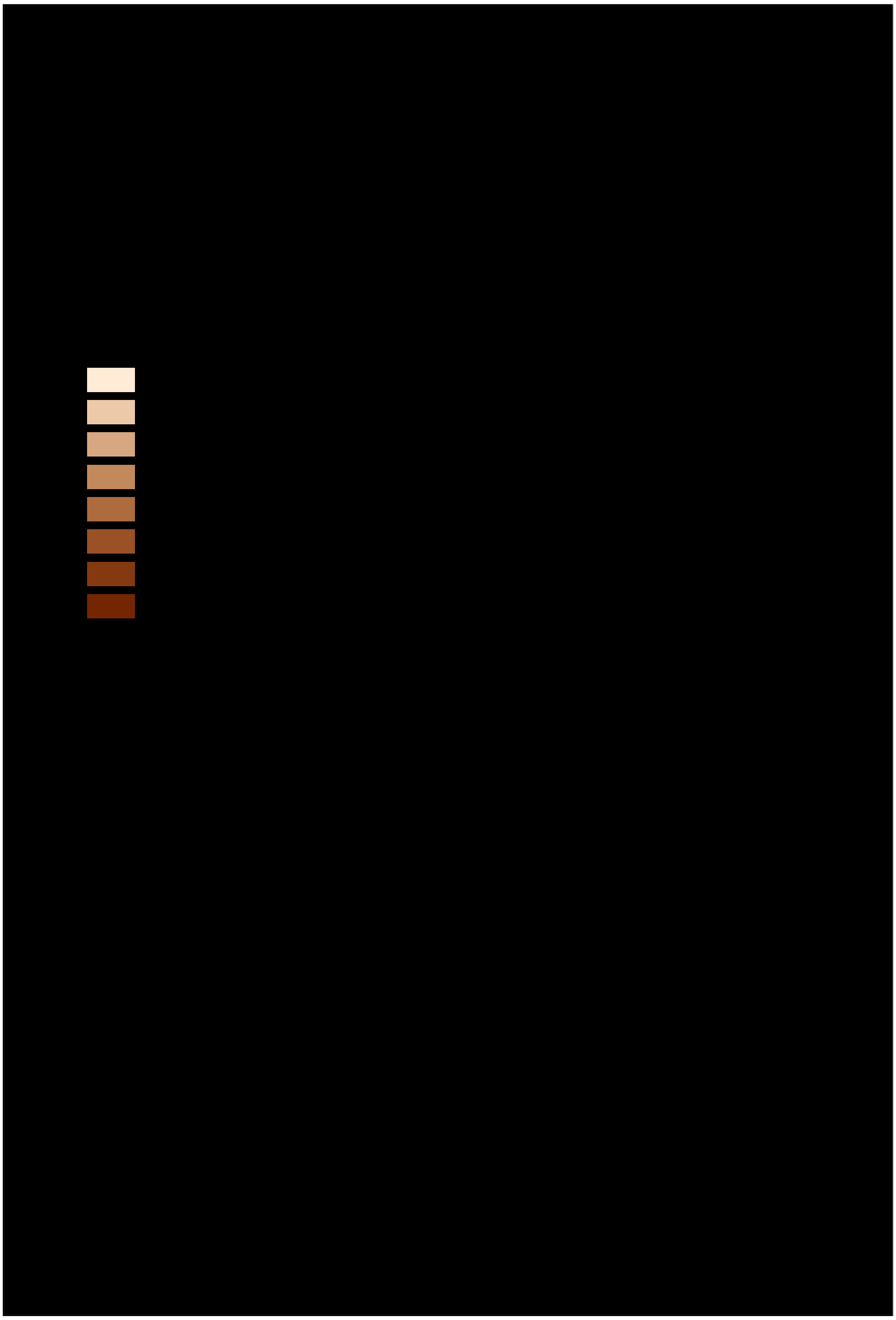

Fig. 6: Solid geology class distribution in England and Wales. Classes from 0 to 8 represent solid geologies with increasing average CSI value for surveyed RHS sites. Low CSI values correspond to fine sediment dominated streams, high CSI to coarse sediment dominated streams.

3.3. Regression kriging

Following data quality checks, 9473 British RHS sites were retained for the analyses (Fig. 7A). The best model, following selection, included the four PCA axes, two solid geology categories, geological age and four hydrometric area groups (Table 2). The model explained 67% of the variability in CSI.

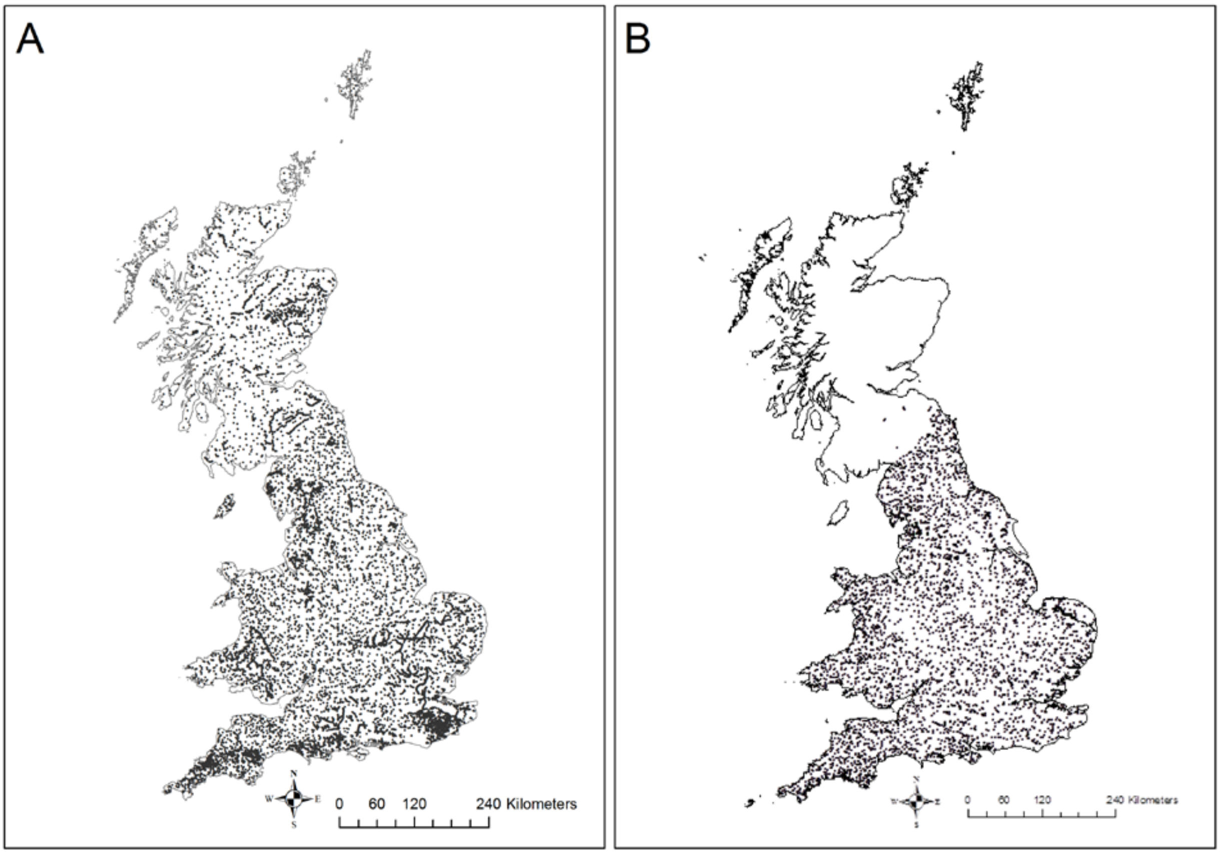

Fig. 7 : Distribution of A) 9473 RHS sites used for modelling CSI and B) 3884 sites used for testing.

The best fit for modelling the OLS and GLS residual variograms was obtained using a combination of spherical and exponential functions (Fig. 8). Spatial correlations were observable up to 5 km and started to plateau after 13 km. The presence of a nugget can be explained by between-surveyor variability as well as time of survey, flow condition etc. The residuals showed a slight tendency for under-prediction, but no great departure from normality (Fig. 9A). Kriged residuals were added to the cross-validated linear model predictions and the estimated R^2 for the spatially corrected model was 0.74. The kriged model residuals showed a marked improvement in prediction with a tighter and more symmetrical distribution around the mean and figures very close to zero (Fig. 9B). A variogram plot of residuals following kriging showed no sign of remaining spatial correlation. The kriged model residuals were also investigated for different years of survey to check the validity of the model over time. The residuals showed no clear pattern of change in distribution between the years of survey with most variability explained by differences in sample size.

Table 2: OLS model linking transformed CSI values to GIS map-derived covariates following a best subset selection procedure.

Model Summary

Model	R	R^2	Adjusted R^2	RMSE
1	0.818	0.670	0.669	1.103

	Sum of Squares	df	Mean Square	F	p
Regression	23345	11	2122.265	1743	< .001
Residual	11522	9466	1.217		
Total	34867	9477			

	Unstandardised	Standard Error	Standardised	t-value	p
intercept	5.978	0.029		209.358	< .001
PCA1	0.566	0.011	0.425	53.725	< .001
PCA2	0.443	0.011	0.262	39.218	< .001
PCA3	0.158	0.021	0.048	7.487	< .001
PCA4	0.062	0.028	0.015	2.240	0.025
Geological age	-0.886	0.048	-0.231	-18.386	< .001
Solid Geology 2	0.301	0.038	0.073	7.831	< .001
Solid Geology 3	0.585	0.072	0.053	8.094	< .001
Hydrometric group 1	-1.525	0.072	-0.158	-21.032	< .001
Hydrometric group 2	-1.073	0.046	-0.272	-23.284	< .001
Hydrometric group 3	-0.558	0.062	-0.068	-9.075	< .001
Hydrometric group 4	-0.540	0.032	-0.124	-16.712	< .001

The model was then tested on 3884 independent sites surveyed in 2006-11 mainly in England and Wales (Fig. 7B). The model explained 64% of the variability in the new data. The residual distribution was centred around zero and showed no tendency for over- or under-prediction (Fig. 9C). Histograms of residuals per year of survey showed no significant pattern.

The model and test data were then joined and the regression kriging model was applied to the entire river network. The resulting map (Fig. 10) shows a clear gradient between the uplands in the West and North of England and Wales dominated by harder, older geologies and coarser substrate types, and the East and South, where sedimentary rocks predominate and channels are dominated by finer sediments.

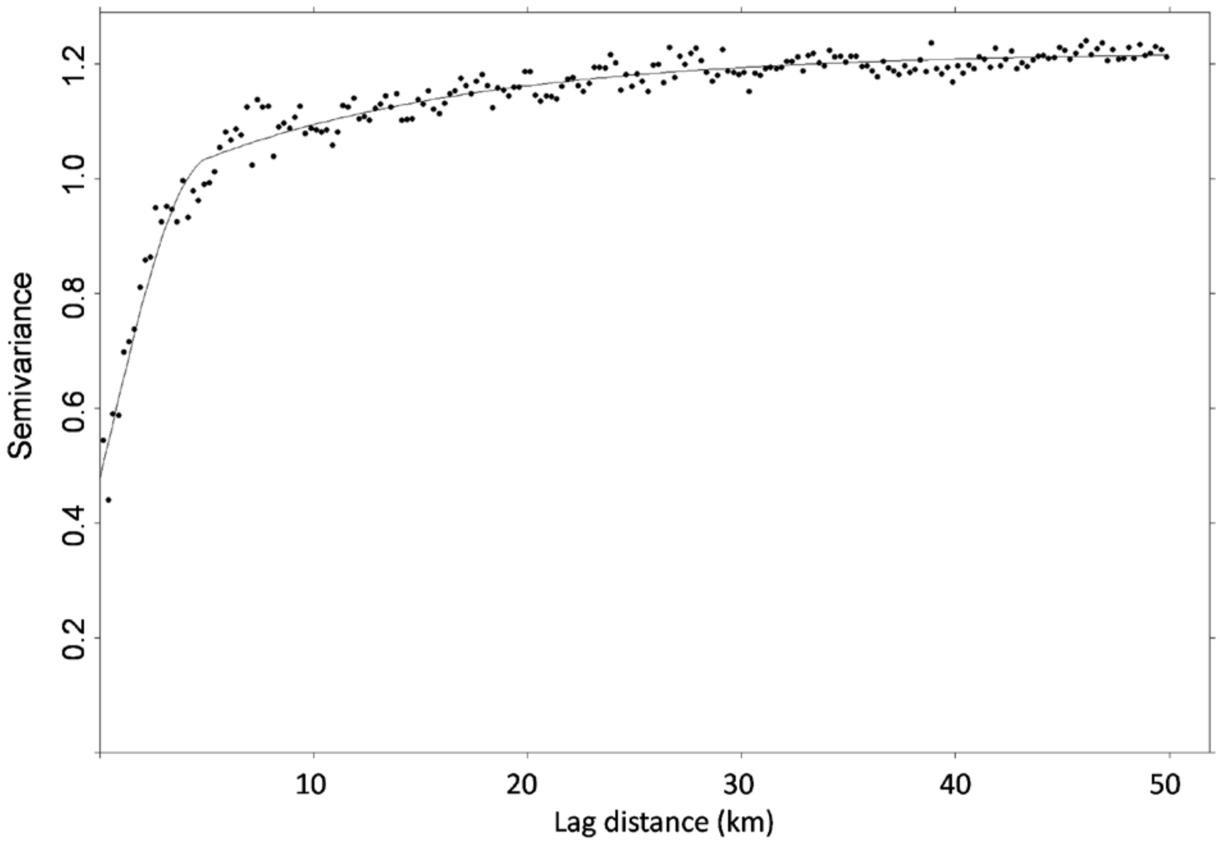

Fig. 8: OLS model residual variogram fitted with a combination of exponential and spherical functions.

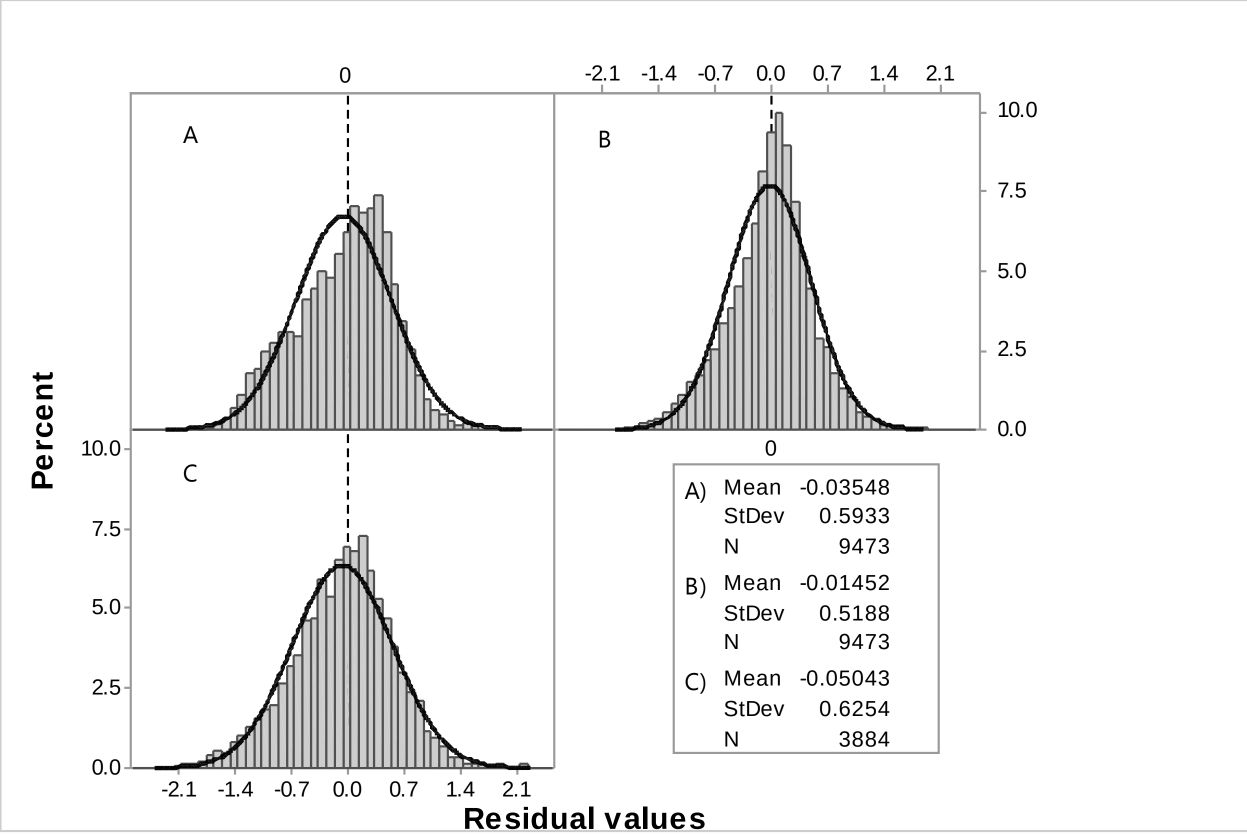

Fig. 9: Distribution of model residual values with fitted curves for the cross-validated GLS linear regression model (A) before and (B) after kriging; and (C) for the test sample.

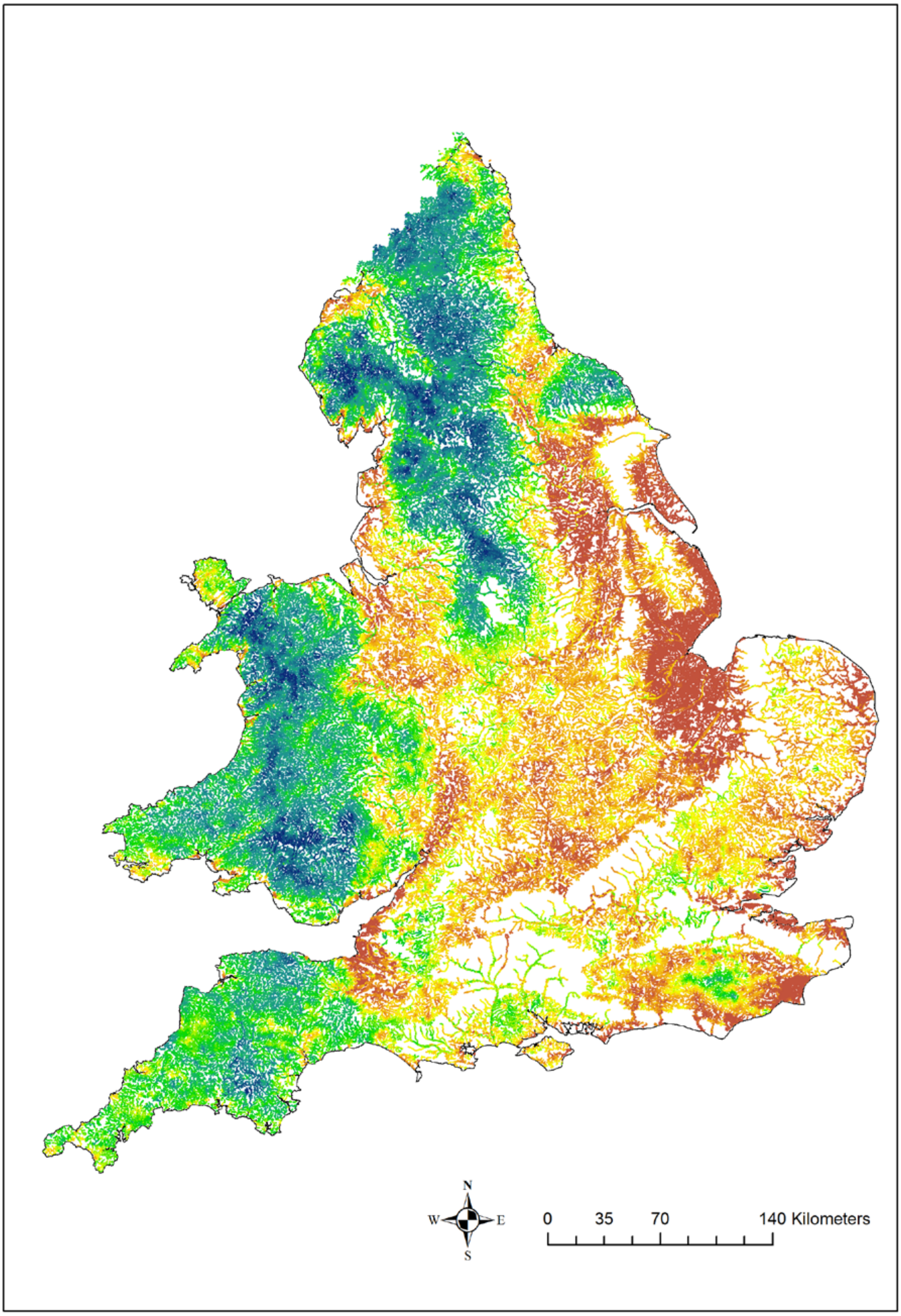

Fig. 10: Map of predicted values of CSI using regression kriging at every 500m across the 1:50,000 river network on a gradient from bedrock/boulder (blue) to gravel-pebble (green)

and silt-sand-clay (brown). White reflects areas of low drainage density where fewer streams are present.

4. Discussion

Using existing data and geostatistical modelling techniques, it was possible to identify and predict channel substrate and apply the model to the entire river network in England and Wales thus, providing environmental practitioners and managers with the first comprehensive national scale map of channel substrate distribution across the network.

Traditionally, substrate is characterised using some quantification of sediment size distribution; generally statistics taken from a distribution such as D_{50} or D_{84} (median size or 84th percentile size). Sediment sizing is based on field survey where substrate is either sampled manually or using mechanical techniques (Kondolf et al., 2003). Sampling efforts can be intensive and, therefore, expensive (Bunte et al., 2009) and there are no existing national datasets of substrate size available at present. Davenport et al (2004) attempted to derive estimates of substrate size using RHS data. The Sediment Calibre Index (SCI) was calculated by multiplying substrate occurrence by Wentworth category median size in phi units and averaging over 10 spot-checks. We produced a modified version of the SCI, the SCI_m by adding one category including bedrock and artificial substrate using the following equation and correcting some of the mistakes that were introduced in the original publication for sand and silt phi values (Angela Gurnell, pers. comm.):

$$SCI_m = (-12(AR+BE) - 8\ BO - 7\ CO - 3.5\ GP + 1.5\ SA + 6\ SI + 9\ CL) / N_{sc}$$

The SCI_m was applied to 10,135 RHS British sites and compared to the CSI. The correlation between the two indices was extremely high (Pearson correlation coefficient = -0.985, n=10,135) suggesting that the CSI and the SCI_m both represent average channel substrate size. It is important to note that CSI and SCI_m are unlikely to represent D_{50}

unless the substrate size distribution follows a unimodal symmetrical distribution. Further studies involving comparisons with more traditional sediment sizing techniques and RHS would help identify the relevance and significance of the CSI with regards to characterising channel substrate across a 500 m reach but these data currently do not exist.

One strong advantage of using multivariate techniques is in the identification of major patterns and dimensions that can be related to biological gradients of species distribution. The assumption is that species distribution and community composition adapt to dominant habitat gradients. In the present case, the main dimension extracted was a well-known substrate fining gradient (Werritty, 1992). Another advantage is that dominant gradients are likely to be influenced by drivers of geomorphological change and are, therefore, more predictable. CSI was correlated and explained by a series of attributes acting at different scales that can be related to known drivers of geomorphological change. At the local (site) scale, Jeffers' PCA1 and PCA3 represent ground slope whilst PCA2 acts as a surrogate for discharge. Slope and discharge are the main drivers of stream power which is strongly related to sediment transport and sorting (Rice and Church, 1998). Solid geology classes and PCA4 provide a wider catchment scale context of geomorphological influence on stream energy and sediment supply. The geology categories reflect the age and hardness of geological types whilst PCA4 represents upstream catchment slope (Jeffers, 1998). Wider scale influences were represented by attributes such as geological age and hydrometric area groups that provided a greater spatial and climatic context for predicting channel substrate.

The geostatistical analysis of model residuals revealed the presence of remaining unexplained spatially correlated variance with a relatively short range. This suggests the presence of local random factors influencing substrate distribution, potentially linked to sediment transport, sediment supply from the surrounding landscape (Church, 2002), riparian land use or human-made impacts. They could also represent non-linear spatial responses to geomorphic drivers or local discontinuities in substrate caused by confluences, landslides/bank erosion or the presence of lakes or reservoirs (Rice et al., 2001).

The kriging process greatly reduced the spatial correlation in the residuals and increased the model predictive power. The model was also tested for its ability to predict substrate for different surveys. Channel substrate and geomorphological forms tend to be quite stable over decadal timescale and evolve slowly unless significant changes occur such as channel modification or catastrophic events (Knighton, 1998). Therefore, we expected to see no large decrease in predictive power across the years of survey. This was confirmed for both the modelling and the test samples which showed very little deviation in predictive power across survey years.

The overall predictive power of the model on the test sample was satisfactory with 64% of the site variability explained by the model. An examination of model residuals for the 100 sites with the largest residual values showed no discernible patterns apart from under-prediction of artificial, bedrock (22 % of the sites) and peat substrates (2% of sites). Other sources of error were investigated. For the 2007-8 RHS baseline survey, one third of sites were surveyed on parts of the 1:50,000 river network not covered by the previous sampling strategy which was based on the 1:250,000 river network. A comparison of residual values between sites located on the 1:250,000 and 1:50,000 networks using ANOVA showed a significant difference between sample means (F=384; p<0.001; n=2772). The model tended to predict slightly coarser substrate size on the 1:50,000 sample than observed. This could be linked to stream size and management regime. Sites selected on the 1:50,000 sample tended to be narrower, with an average bankfull width of 4.1 m (*n*=1692), compared to 9 m (*n*=3134) for the 1:250,000 sites, with a strong presence of agricultural ditches and artificial channels. It is possible that the model parameters do not fully account for small artificial channels and this shows some of the limitations of using the predictive model on sites collected at different scales. The overall impact of scale on predictive accuracy was, however, small and produced only a slight decrease in overall predictive accuracy. A practical advantage of RK is that it honours field data. Thus, predictions using RK will always fit perfectly the observed values at surveyed sites. This is important from a predictive accuracy viewpoint, but also on an operational viewpoint as it reinforces the credibility of the model in the eyes of users (Naura, 2014).

5. Conclusion

We proposed an alternative approach for mapping habitat elements across *entire* river networks that makes use of existing semi-quantitative survey data, GIS map-based covariate data and RK. A new national scale substrate index has been developed, which is accurate from 500m up to national scales. This application shows the potential power of using spatially explicit techniques for modelling river attributes at the national scale. The analyses presented in this article are part of a broader effort to characterise and map river habitats, identify river reaches for environmental management and develop practical tools for impact assessment, diagnostics and management planning that will be demonstrated in subsequent publications.

Acknowledgements

RHS data can be accessed through the UK government data portal (http://data.gov.uk/).The research was partly funded through grants provided by the School of Geography at the University of Southampton, the Environment Agency, the

Scottish Environment Protection Agency and the EPSRC. The Authors would like to dedicate this work in memory of Prof Mike Clark, whose dedication to the science and practice of environmental and specifically river management is greatly missed.

References

Bivand, R.S., Pebesma, E.J., Gomez-Rubio, V., 2008. Applied spatial data analysis with R. Springer, New York.

Bunte, K., Abt, S.R., Potyondy, J.P., Swingle, K.W., 2009. Comparison of Three Pebble Count Protocols (EMAP, PIBO, and SFT) in Two Mountain Gravel-Bed Streams. JAWRA Journal of the American Water Resources Association 45, 1209-1227.

CEN, 2004. A guidance standard for assessing the hydromorphological features of rivers., in: Comité Européen de Normalisation (Ed.).

Cherrill, A., Mcclean, C., 1999. Between-observer variation in the application of a standard method of habitat mapping by environmental consultants in the UK. Journal of Applied Ecology 36, 989-1008.

Chessman, B.C., Fryirs, K.A., Brierley, G.J., 2006. Linking geomorphic character, behaviour and condition to fluvial biodiversity: implications for river management. Aquatic Conservation-Marine and Freshwater Ecosystems 16, 267-288.

Church, M., 2002. Geomorphic thresholds in riverine landscapes. Freshwater Biology 47, 541-557.

Collen, B., Whitton, F., Dyer, E., Baillie, J., Cumberlidge, N., Darwall, W., Pollock, C., Richman, N., Soulsby, A., Bohm, M., 2014. Global patterns of freshwater species diversity, threat and endemism. Global Ecology and Biogeography 23, 40-51.

Collins, A.L., Jones, J.I., Sear, D.A., Naden, P.S., Skirvin, D., Zhang, Y.S., Gooday, R., Murphy, J., Lee, D., Pattison, I., Foster, I.D.L., Williams, L.J., Arnold, A., Blackburn, J.H., Duerdoth, C.P., Hawczak, A., Pretty, J.L., Hulin, A., Marius, M.S.T., Smallman, D., Stringfellow, A., Kemp, P., Hornby, D., Hill, C.T., Naura, M., Brassington, J., 2014. Extending the evidence base on the ecological impacts of fine sediment and developing a framework for targeting mitigation of agricultural sediment losses. Defra report WQ128, Defra ed. Defra.

Davenport, A.J., Gurnell, A.M., Armitage, P.D., 2004. Habitat survey and classification of urban rivers. River Research and Applications 20, 687-704.

Downs, P.W., 1995. River channel classification for channel management purposes. Changing River Channels. John Whiley & Sons Ltd., University of Nottingham, UK, 347-367.

Environment Agency, 2012. National Risk Assessments for Morphology: Broad-Scale Evaluation. Environment Agency, Bristol, p. 31.

European Environment Agency, 2012. European waters - assessment of status and pressures, p. 100.

European Union, 2000. Directive 2000/60/EC of the European Parliament and of the Council of 23 October 2000 establishing a framework for Community action in the field of water policy.

Fuller, R.M., Smith, G.M., Sanderson, J.M., Hill, R.A., Thomson, A.G., 2002. The UK Land Cover Map 2000: Construction of a parcel-based vector map from satellite images. Cartographic Journal 39, 15-25.

Gasparini, N.M., Tucker, G.E., Bras, R.L., 1999. Downstream fining through selective particle sorting in an equilibrium drainage network. Geology 27, 1079-1082.

Greenacre, M.J., 1993. Correspondence analysis in practice. Academic Press Limited, San Diego.

Hendry, K., Cragg-Hine, D., 1997. Fisheries Walkover Surveys. A Guidance Manual, Fisheries Techincal Manual, Bristol

Hornby, D.D., 2010. RivEX, 6.7 ed. http://www.rivex.co.uk.

Jeffers, J.N.R., 1998. An ordination of river habitats using RHS data. Aquatic Conservation: Marine and Freshwater Ecosystems. 8, 529-540.

Knighton, D., 1998. Fluvial forms and processes - A new perspective. Arnold, London.

Kondolf, G.M., Lisle, T.E., Wolman, G.M., 2003. Bed sediment measurement, in: Kondolf, G.M., Piégay, H. (Eds.), Tools in Fluvial Geomorphology. John Wiley & Sons, pp. 347-395.

Loh, J., Green, R., Ricketts, T., Lamoreux, J., Jenkins, M., Kapos, V., Randers, J., 2005. The Living Planet Index: using species population time series to track trends in biodiversity. Philosophical Transactions of the Royal Society B-Biological Sciences 360, 289-295.

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

Minitab, I., 2010. Minitab 16 Statistical Software, 16 ed. State College.

Montgomery, D.R., Buffington, J.M., 1997. Channel-reach morphology in mountain drainage basins. Geological Society of America Bulletin 109, 596-611.

Morris, P.H., Williams, D.J., 1999. A worldwide correlation for exponential bed particle size variation in subaerial aqueous flows. Earth Surface Processes and Landforms 24, 835-847.

Mosley, M.P., 1987. The classification and characterization of rivers, in: Richard, K.S. (Ed.), River Channels: Environment and Process. Blackwell, Oxford, pp. 295-320.

Naiman, R.J., Beechie, T.J., Benda, L.E., Berg, D.R., Bisson, P.A., MacDonald, L.H., O'Connor, M.D., Olson, P.L., Steel, E.A., 1992. Fundamental elements of ecologically healthy watersheds in the Pacific Northwest coastal ecoregion, Watershed management. Springer, pp. 127-188.

Naura, M., 2014. Decisions Support Systems. Factors affecting their design and implementation within organisations. Lessons from two case studies. Lambert Academic Publishing, Berlin.

Newson, M.D., Clark, M.J., Sear, D.A., Brookes, A., 1998. The geomorphological basis for classifying rivers. Aquatic Conservation-Marine and Freshwater Ecosystems 8, 415-430.

Raven, P.J., Fox, P., Everard, M., Holmes, N.T.H., Dawson, F.H., 1997. River habitat survey: A new system for classifying rivers according to their habitat quality. Freshwater Quality: Defining the Indefinable?, 215-234.

Revenga, C., Campbell, I., Abell, R., de Villiers, P., Bryer, M., 2005. Prospects for monitoring freshwater ecosystems towards the 2010 targets. Philosophical Transactions of the Royal Society B-Biological Sciences 360, 397-413.

Rice, S., Church, M., 1998. Grain size along two gravel-bed rivers: Statistical variation, spatial pattern and sedimentary links. Earth Surface Processes and Landforms 23, 345-363.

Rice, S.P., Greenwood, M.T., Joyce, C.B., 2001. Tributaries, sediment sources, and the longitudinal organisation of macroinvertebrate fauna along river systems. Canadian Journal of Fisheries and Aquatic Sciences 58, 824-840.

Rosgen, D.L., 1994. A Classification of Natural Rivers. Catena 22, 169-199.

Sala, O., Chapin, F., Armesto, J., Berlow, E., Bloomfield, J., Dirzo, R., Huber-Sanwald, E., Huenneke, L., Jackson, R., Kinzig, A., Leemans, R., Lodge, D., Mooney, H., Oesterheld, M., Poff, N., Sykes, M., Walker, B., Walker, M., Wall, D., 2000. Biodiversity - Global biodiversity scenarios for the year 2100. Science 287, 1770-1774.

Sear, D., Newson, M., Hill, C., Old, J., Branson, J., 2009. A method for applying fluvial geomorphology in support of catchment-scale river restoration planning. Aquatic Conservation-Marine and Freshwater Ecosystems 19, 506-519.

Strayer, D., Dudgeon, D., 2010. Freshwater biodiversity conservation: recent progress and future challenges. Journal of the North American Benthological Society 29, 344-358.

Thorne, C.R., 1997. Channel types and morphological classification. Applied fluvial geomorphology for river engineering and management, 175-222.

Tockner, K., Stanford, J.A., 2002. Riverine floodplains: present state and future trend. Environmental Conservation 29, 308-330.

Townsend, C.R., Hildrew, A.G., 1994. Species traits in relation to a habitat templet for river systems. Freshwater Biology 31, 265-275.

Vaughan, I., Merrix-Jones, F., Constantine, J., 2013. Successful predictions of river characteristics across England and Wales based on ordination. Geomorphology 194, 121-131.

Vaughan, I.P., Diamond, M., Gurnell, A.M., Hall, K.A., Jenkins, A., Milner, N.J., Naylor, L.A., Sear, D.A., Woodward, G., Ormerod, S.J., 2009. Integrating ecology with hydromorphology: a priority for river science and management. Aquatic Conservation-Marine and Freshwater Ecosystems 19, 113-125.

Vaughan, I.P., Ormerod, S.J., 2003. Improving the quality of distribution models for conservation by addressing shortcomings in the field collection of training data. Conservation Biology 17, 1601-1611.

Vorosmarty, C.J., McIntyre, P.B., Gessner, M.O., Dudgeon, D., Prusevich, A., Green, P., Glidden, S., Bunn, S.E., Sullivan, C.A., Liermann, C.R., Davies, P.M., 2010. Global threats to human water security and river biodiversity. 467, 555-561.

Webster, R., Oliver, M.A., 2007. Geostatistics for environmental scientists, 2nd edition ed. John Wiley & Sons Ltd, Chichester.

Wentworth, C.K., 1922. A scale of grade and class terms for clastic sediments. The Journal of Geology, 377-392.

Werritty, A., 1992. Downstream fining in a gravel bed river in Southern Poland: Lithological controls and the role of abrasion, in: P., B., R.D., H., C.R., T., P., T. (Eds.), Dynamics of Gravel Bed Rivers. Wiley: Chichester, pp. 333–346.

World Conservation Monitoring Centre, 1998. Freshwater Biodiversity: a preliminary global assessment, WCMC Biodiversity Series. World Conservation Monitoring Centre, Cambridge, UK, p. 132.

WWF, 2014. Living Planet Report 2014. Species and spaces, people and places. WWF, p. 176.