Testing and Further Development of RIVPACS

Stage 4 Report

An evaluation of procedures for acquiring environmental variables for use in RIVPACS from a GIS

R&D Technical Report E1-007/TR1

D D Hornby, R T Clarke, J F Wright and F H Dawson

Research Contractor
CEH Dorset

Centre for Ecology & Hydrology
NATURAL ENVIRONMENT RESEARCH COUNCIL

ENVIRONMENT AGENCY

Publishing Organisation
Environment Agency, Rio House, Waterside Drive, Aztec West, Almondsbury, Bristol BS32 4UD
Tel: 01454 624400 Fax: 01454 624409
Website: www.environment-agency.gov.uk

ISBN: 1 8570 58445

Dissemination Status
Internal: Released to Regions
External: Released to Public Domain

Statement of Use
This report is part of the output from a project to test and further develop RIVPACS (a computer programme which predicts the river macroinvertebrate communities to be expected at high quality sites). It describes the use of a Geographic Information System (GIS) to derive the existing RIVPACS environmental variables, plus some new variables, for each of the RIVPACS reference sites (and also Agency test sites, if required). It compares the ability of the original and GIS derived data to place sites in their correct RIVPACS classification group (a measure of their effectiveness in the prediction process) and of the new variables to improve this process.

The report is to be used by the Agency to guide future development of RIVPACS. It will also influence development of Agency biology data systems and their links to GIS.

Keywords
RIVPACS; biological monitoring; macroinvertebrates; environmental variables; GIS

Research Contractor
This document was produced under R&D Project E1-007 by :
CEH Dorset, Winfrith Technology Centre, Winfrith Newburgh, Dorchester, DT2 8ZD
Tel : 01305 213500 Fax : 01305 213600

Environment Agency Project Manager
The Environment Agency's Project Manager for R&D Project E1-007 was:
Dr R A Dines, Southern Region

Further copies of this report are available from:
Environment Agency R&D Dissemination Centre
WRc, Frankland Road, Swindon, Wilts. SN5 8YF
Tel: 01793 865000 Fax: 01793 514562 E-mail: publications@wrcplc.co.uk

ACKNOWLEDGEMENTS

This research project was funded by the Environment Agency. We would like to thank the Agency for continuing to support the development of RIVPACS. In particular we are grateful for the help and guidance provided by the Agency's designated Project Manager, Dr R A Dines and R&D Management Coordinator, Ms P Mardon. Thanks also to Mr B Hemsley-Flint, who managed the early stages of the project and to Dr J Murray-Bligh for the initial specification.

During the project, it became apparent that the amount of development work required in order to extract reliable information from the GIS in an efficient manner was much greater than originally envisaged. However, it was also clear that this research would have practical application to the River Habitat Survey (RHS). As a result, additional funding from the Agency was made available to Dr H Dawson to enable development work to encompass a wider range of variables throughout the river network of Great Britain.

The original blue line network was supplied by CEH Wallingford (formerly the Institute of Hydrology, Wallingford). We are grateful to CEH Wallingford and in particular to Mr D Morris and Ms S Renn for the provision of advice and additional data at various stages during the project.

Some preliminary work on the blue line network was undertaken by Dr G Irons, under the supervision of Dr J Hilton at the IFE River Laboratory. Mr D Hornby then took over and has been responsible for undertaking the enormous task of checking, editing and processing the GIS network throughout Great Britain.

EXECUTIVE SUMMARY

RIVPACS requires information on a small suite of environmental variables in order to make predictions of the macroinvertebrate fauna to be expected at a given site with stated environmental features in the absence of environmental stress. The Option 1 suite of variables has remained the same in RIVPACS II, (available in 1990) and in RIVPACS III/RIVPACS III+ (used in the 1995/2000 General Quality Assessments).

Some of the current variables including slope of site, altitude of site and distance of site from source, together with mean annual discharge category are acquired manually from maps. This is a time-consuming process which can be prone to error.

In this package, we have examined the feasibility of developing procedures for acquiring accurate values of these predictor variables from a Geographic Information System (GIS). We have also investigated whether additional site variables acquired from the GIS are capable of increasing the accuracy of RIVPACS predictions.

The software ArcView, a Windows-based GIS, was used in conjunction with the 'blue line' river system of Great Britain, as shown on 1:50000 scale Ordnance Survey Landranger maps to create a novel procedure for the extraction of environmental data. This involved two separate processes, each one of which was very time-consuming.

First, it was necessary to check and edit a variety of errors in the original digitised blue-line network for each individual hydrometric area throughout Great Britain. Second, the decision was taken to 'process' the entire network by incorporating information on the routes to the source and mouth of each river from any given arc on the network. This information is essential for calculating such features as distance to source and slope. By undertaking this mammoth task prior to routine use of the purpose-built system for extracting site information, the query process to automate estimation of the values of the various variables became much more efficient.

Some problems remain. At present, site grid references are only given to 100m resolution; this is not always sufficiently accurate to automatically associate the site with the correct river location on the blue-line network. Therefore each new biological site needs to be manually assigned to the correct river location on the network before the GIS can be used to derive environmental data for the site. In a small number of low-lying areas of the country, the grid-like drainage pattern can not currently be resolved by the GIS in order to calculate distance to source, slope etc. Hence environmental variables at 49 of the 614 RIVPACS reference sites were unobtainable using GIS, and the original map-derived RIVPACS values for these variables were used to maintain a full dataset in subsequent multiple discriminant analyses (MDA).

Values of current RIVPACS predictor variables acquired from the GIS and from maps were compared for the RIVPACS reference sites. Careful manual re-checking found that the use of GIS to obtain altitude and distance from source was generally more reliable, in addition to being quicker. Different procedures, often involving different distances up- and down- stream, were used to obtain slope of site by the GIS and the manual map method, resulting in a limited number of substantial changes in slope estimates. There were further practical problems in obtaining discharge category from the GIS, because this layer was supplied at a different scale to the blue line and it had to

be loaded on top of the blue line network before manually selecting the site and hence the discharge category. However, over 90% of sites were within one discharge category of the original map-read category.

The GIS-derived values of these four variables were then used with the remaining RIVPACS variables in MDA to determine whether, overall, they gave improved ability to predict the biological group of each of the 614 RIVPACS reference sites. The GIS-derived values are assumed to be estimated with greater accuracy, but they failed to improve on current ability to predict the biological group of the sites.

GIS software procedures were developed to derive new environmental predictor variables for any river site. These were: altitude of source, slope to source, a measure of stream power, upstream catchment area, the proportion of the upstream catchment covered by each major RHS solid and drift geology class and the solid and drift geology class of the 1 km square containing the site. When assessed on the RIVPACS reference sites, none of these GIS variables provided any significant improvement in ability to predict site group and hence expected fauna when used with the existing RIVPACS predictor variables. There was only some minor improvement when adding stream power (a function of discharge, stream width and slope at site).

However, the automated GIS procedures identify an incorrect upstream catchment for a small percentage of sites (5-10%); developing GIS procedures to "manually" correct such site positions is beyond the scope of this project, but merits further investigation.

In the long-term, it would appear that the use of GIS-derived values of some RIVPACS predictor variables will bring benefits in terms of speed of acquisition, greater accuracy and a modest increase in predictive capability. However, the consequences of implementing such a change on the RIVPACS software itself and on all Ecological Quality Index (EQI) and other results generated by RIVPACS are such that it should be carried out as a single operation and only when there is clear evidence that progress in generating reliable outputs from the GIS is complete.

We also propose that there is a need for a Windows version of RIVPACS before additional changes are made to the current software.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1 Background

At the outset, it should be pointed out that the original title of this package was 'An evaluation of new environmental variables for RIVPACS'. However, in the early stages of the work, it became apparent that RIVPACS (River InVertebrate Prediction And Classification System) would become a more efficient management tool if some of the *existing* variables used for prediction could be acquired in an objective manner from a Geographic Information System (GIS). Development of this approach would also provide opportunities for acquiring additional variables useful for prediction of the macroinvertebrate fauna. Hence, the focus of this package has changed over time, to incorporate the original concept but also to seek to improve the capture of existing variables.

RIVPACS III+ requires information on a small suite of environmental variables in order to make predictions on the macroinvertebrate fauna to be expected at a given site with stated environmental features in the absence of environmental stress. However, during the early stages of the development of RIVPACS, around 30 environmental variables were used for prediction, before smaller sub-sets of variables were found to be capable of generating acceptable predictions of the expected fauna (Wright, 2000). Hence, a considerable amount of research has already gone into consideration of those variables expected to have predictive capability.

RIVPACS III+ offers five separate environmental options for prediction of the fauna, although Option 1 is recommended for use in Great Britain (Table 1.1). As is apparent from the table, eight variables are used in all five options and further variables are also needed which vary with the option chosen.

Table 1.1: The five environmental options available for prediction in RIVPACS III+

All options require the following eight variables:

Distance from source (km)	Discharge category (9 groups, cumecs)
Altitude (m)	Mean water width (m)
Latitude (^{o}N)	Mean water depth (cm)
Longitude (^{o}W)	Mean substratum (phi units)

Some additional variables are also required, according to the option chosen:

Option	1	2	3	4	5
Alkalinity (mg $CaCO_3$ l^{-1})	+	+	-	+	-
Slope (m km^{-1})	+	-	+	+	-
Mean air temperature (^{o} C)	+	+	+	-	+
Air temperature range (^{o} C)	+	+	+	-	+

The current variables used for prediction are acquired from four separate sources:

1. Variables obtained manually from Ordnance Survey (OS) 1:50,000 scale Landranger Maps
   National Grid Reference (NGR) (see 2 below)
   Slope of site
   Altitude of site
   Distance of site from source

2. Data derived internally by RIVPACS from the National Grid Reference
   Latitude
   Longitude
   Mean air temperature
   Air temperature range

3. Information provided by Environment Agency staff
   Mean annual discharge for the site, expressed as one of ten categories (1-10)
   Total alkalinity at the site (mean of all determinations for one year)

4. Field data based on measurements made at the site in spring, summer and autumn
   Mean stream width
   Mean stream depth
   Mean substratum

Further information on the detailed procedures used to acquire each of these four sources of data are given in the Environment Agency procedures manual (Murray-Bligh *et al.*1997) and will not be repeated here.

From the listing of variables given above, it is apparent that sources 1, 2 and 3 are essentially fixed values for the site (although mean alkalinity may vary slightly from year to year). Hence, once collected, they are available for use in future years without further effort. In contrast, the field-based variables are expensive to collect because of the requirement for three visits, and may vary from year to year, particularly if extremes of weather are encountered on site. For some time, we have held the view that there would be merit in obtaining long-term mean values of those variables subject to inter-annual changes, in order to provide a long-term fixed prediction of the fauna to be expected at a given site. This would mean that the E value in O/E ratios would remain constant between successive quinquennial General Quality Assessment (GQA) surveys and all changes in grading would be due to changes in the observed (O) fauna. An alternative strategy is to look for alternatives to the predictor variables listed under source 4. It would be essential for these variables to be fixed values for the site and not subject to inter-annual change.

This is an ideal time to re-examine the range of variables used for prediction and the procedures used to acquire these variables. Although the map-based variables are not subject to changes over time, they take time to abstract and may be measured with some error. Hence, there would be many practical benefits from acquiring a number of the existing, and also some new environmental variables from a Geographic Information System, in order to make the estimation process quick, objective and repeatable.

1.2 Objectives

The overall objective is as follows:

'To produce a scoping report on the potential for acquiring some of the existing and also some new environmental variables for prediction from a Geographic Information System and its associated environmental databases'.

The specific objectives (modified from the original specification with the agreement of the nominated officer) are as follows:

1. To investigate the feasibility of developing procedures for acquiring existing RIVPACS predictor variables from a GIS.

2. To attempt to acquire new RIVPACS predictor variables from a GIS.

3. To assess the merits of existing and new variables derived from a GIS in RIVPACS predictions.

4. To make recommendations based on item 3, regarding the desirability of acquiring variables for use in prediction through a GIS. The procedures should be described and outline costs should be provided.

2. ACQUIRING THE GIS VARIABLES

2.1 Introduction to ArcView and the Spatial Data Available

The software ArcView is a customisable Windows-based Geographical Information System (GIS) that allows the user to display and query spatial data. Spatial data is information that has co-ordinates and attributes. For example, a point has XY coordinates and its attribute may label it as a house or an oil well. A line is composed of a series of points but may represent a road, river or power line. In view of the fact that a number of specialist terms are used in this report, a brief glossary is provided (Table 2.1).

Table 2.1 Glossary of GIS terms

ARC	An arc is a line. Arcs have direction; they start FROM a point and go TO a point. Arcs are themselves a series of points (vertices). To maintain topology, arcs are only allowed to join each other at their end points (nodes). They are not allowed to cross themselves or other arcs within the same coverage.
COVERAGE	A layer of information with spatial co-ordinates and attributes. A coverage for rivers should not contain arcs that represent roads. These would be in a separate coverage.
NODES	A node is an end point of an arc. There are two types of node. An arc starts at an *Fnode* and ends at a *Tnode*. Nodes are vertices of an arc.
POINTS	A point is the simplest spatial feature and represents one point in space.
ROUTES	ArcView can search a network of arcs and generate routes between two or more points. A route can be considered as an arc, generated in a separate route layer which lies over the network and starting and ending at the points the user has specified on the network.
SCRIPT	Arcview's programming language, AVENUE, is a script language. Script languages use traditional programming syntax but are able to call upon libraries of optimised code to execute complex tasks.
SHAPEFILE	The file format used to hold the spatial data and its attributes.
SNAPPING	The process by which arcs are forced to join each other at their nodes so network topology can be maintained.
VERTEX	A point within an arc. Two or more vertices are required to form an arc.

The ‘blue line’ network is the river system of Britain as displayed on the 1:50,000 scale Ordnance Survey Landranger map series. This network was created by the former Institute of Hydrology, now the Centre for Ecology and Hydrology (CEH) Wallingford, and was made available to the Institute of Freshwater Ecology (IFE) River Laboratory (now part of CEH Dorset) for research purposes. To make a fully connected river network, CEH Wallingford “centre lined” lakes and lochs by digitising straight lines through the middle of such water bodies. This centre-lining of lakes makes the dataset superior to Ordnance Survey data in that you are able to trace from mouth to source. By maintaining topology (strict rules about the connectivity within spatial data) and generating specific coding within the attributes, the 'blue line' river network can be

queried for spatial information such as altitude at site, slope at site, distance to source etc.

Altitudes were only supplied at points along the course of each river. Heights were supplied in this format due to the prohibitive cost of the CEH Wallingford Digital Terrain Model.

CEH Wallingford also provided a 50m pixel flow direction grid for the whole of Great Britain. This is a grid of cells with integer values that indicate general surface slope direction and can be used, with appropriate GIS commands, to define catchment boundaries.

The 1:625,000 scale British Geological Survey (BGS) information on Drift and Solid Geology were supplied to CEH by the BGS. These spatial layers were transferred from a Unix workstation to PC and incorporated into the GIS system without any pre-processing or editing of the data.

A 1:250,000 scale river network showing discharge categories was supplied by the Environment Agency for England and Wales. Due to the differences in scale between this network and the blueline network (1:50,000), rivers do not overlie one another perfectly; this makes automatic site selection prone to error if discharge is required. Hence, the correct discharge category for a selected site was obtained by manually selecting the appropriate river in the discharge category layer of the integrated GIS.

Initially, much of the data was stored on a SUN workstation at CEH Dorset but it was later transferred to PC in order to develop it within a more accessible user-friendly Windows environment.

A considerable amount of editing and processing of the data has been necessary and these improvements to the GIS datasets are continuing. This is necessary in order that the data can be queried rapidly. During the initial stages of development, the process of finding the source was repeated every time for each site and for small hydrometric areas this was acceptable. However, the same process became too slow (several hours for one site) for larger networks such as the River Thames. Hence, early in 2000, an alternative approach of “hard wiring” the nodes within the network was adopted and as a consequence considerable pre-processing was required. The advantages of this approach are immediately obvious, because a full query on a medium sized catchment such as the River Ribble can take just 20 seconds, returning the following information: co-ordinates of site, hydrometric area, OS map number, Agency public face and water management areas, solid and drift geology, catchment area, total length of upstream water courses, distance to source, coordinates of source, distance to mouth, co-ordinates of mouth, altitude of site, altitude of source and slope at site.

The modular design of the ArcView project allows new variables to be extracted from the data or new layers of information to be queried when they are added to the GIS system.

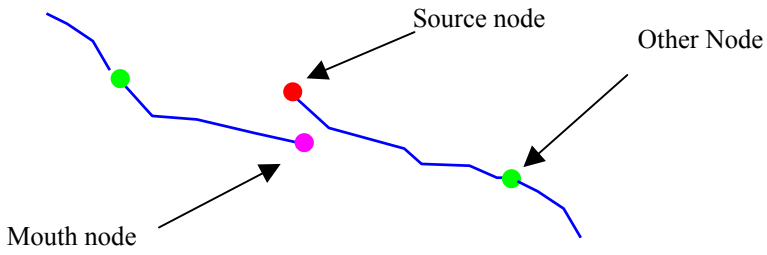
2.2 Initial Checking and Editing of the Blue Line Network

Before the blue line network could be used for tracing and acquiring map variables, it was essential to check the entire network and edit out imperfections. Although the majority of the supplied river network was acceptable there were, nevertheless, a number of imperfections such as breaks in the network. For example, where a river flowed under a road bridge, its blue line representation was interrupted on the printed OS map to show the continuous overlying road; the original blue line network was effectively obtained from a direct image digitisation of the printed OS maps, with all their breaks in the blue lines for rivers. If these break errors were not corrected, it would mean that entire tributaries would be missed out and false results would be returned. This section and section 2.3 of the report detail the full range of imperfections and errors found in the blue line network which were corrected during the editing and subsequent processing phase. Various "pitfalls" which a potential user of the blue line network system needs to bear in mind are also highlighted.

A range of different errors were found and corrected during the editing of the blue line network. These are listed below in sequence from the most straightforward to the most difficult. At the outset, the United Kingdom (UK) was divided into its component Hydrometric Areas and each one was corrected in turn, in order to progress in a systematic and efficient manner.

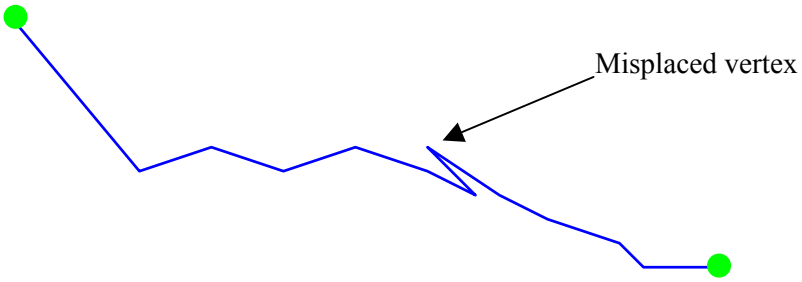
2.2.1 Un-joined arcs

When arcs have not been joined up, the network is incomplete. These imperfections were easy to spot because source nodes (i.e. *Fnode* not joined to any other arcs) were labelled red and mouth nodes (i.e. *Tnode* not joined to any other arcs) were labelled magenta. If a red and a magenta node appeared very close together within the network then it was indicative of a break. In this case, the source and mouth nodes were removed and the arcs were snapped together. Sometimes these breaks were the result of minor mismatches between two adjacent maps at digitisation but in general they were simple breaks.



2.2.2 Misplaced vertex

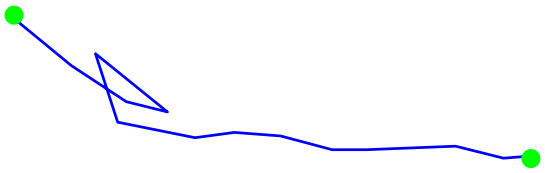
Sometimes, during the digitising process, an arc was "badly" digitised. In this case the resulting arc had a kink in it, which simply did not exist on the map. These very common but minor errors were corrected when they were very obvious.



Note: Throughout all the diagrams in this section, all nodes treated as source nodes are marked as red circles (•), all nodes treated as mouth nodes are marked as magenta (•), and all other nodes are marked as green circles (•).

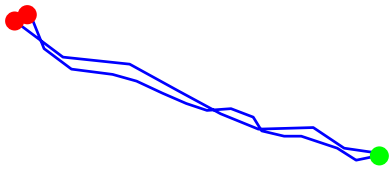
2.2.3 Self crossing arc

Again, during the digitising process, an error was sometimes made such that an arc crossed itself. This could be considered as a more extreme version of the misplaced vertex. As before, vertices were removed to eliminate the crossing over.



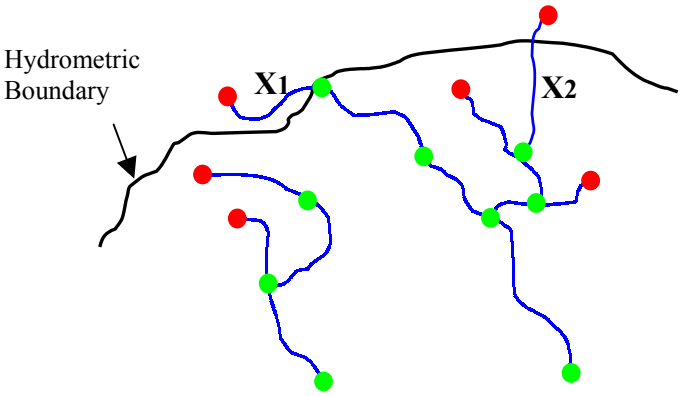
2.2.4 Double digitising

Occasionally, during the digitising process, the arcs were digitised twice. In this case, one of the arcs was removed.



2.2.5 Floating arcs

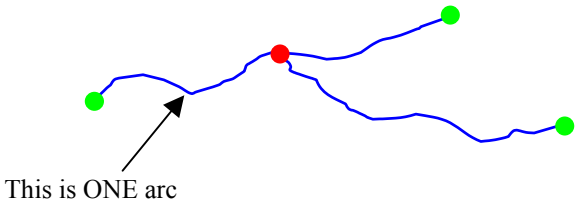
During the division of the river network into separate Hydrometric Areas, arcs representing headwater streams were occasionally left in the wrong Hydrometric Area due to imperfections in the Hydrometric Area boundaries. Hence, the decision was made that if an arc had its *centre* outside the hydrometric boundary under consideration, as for X_1 in the diagram below, then it was removed. In contrast, arcs having their *centre* within the hydrometric boundary, as for X_2, were retained. X_2 type arcs would otherwise end up floating on the edge of the wrong network. In practice, they were cut and pasted into the correct hydrometric network. It was also necessary to move the nodes and height point into their respecting coverages.



X_1 = Centre of ARC. This ARC would be deleted
X_2 = Centre of ARC. This ARC would be retained

2.2.6 Incorrect node connections

These were found when a red source node was within a network but was not a break. The arc whose Fnode was labelled as a source simply ended on top of an arc where there was no node to join them. A break was made, the arcs were snapped together and the source node removed.



2.2.7 Missing or additional arcs

Occasionally, sections of watercourse were missing from the blue line network (although present on the 1:50,000 OS Landranger maps) or, alternatively, were present in the digital layer but absent from the 1:50,000 OS Landranger maps. In two cases, additional arcs turned out to be gridlines; these were removed. In other cases the arcs were canals and these were cut out and pasted into the shapefile which contains the arcs for canals. The remaining cases were arcs that were genuinely parts of streams.

2.2.8 Extreme example of missing arcs

In Hydrometric Area 53, the major river is the Bristol Avon which flows through Bath. An entire tributary, sections of a second tributary and headwater streams of a third tributary were all missing. Fortunately they all appeared within the canal coverage. These arcs were cut from the canal coverage and pasted into the river network for Hydrometric Area 53. Nodes had to be created and given unique ID numbers, snapped to the arc and then the arc attribute table updated with the new node numbers. Nodes that represented stream sources had to be copied and pasted into the source coverage and the source attribute table updated. This was an extremely time consuming task.

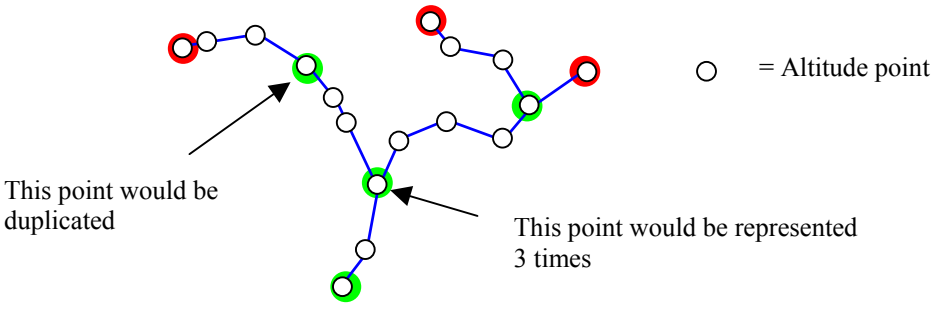
2.2.9 Merging miscellaneous data

A meeting in March 2000 with Mr D Morris of CEH Wallingford provided an explanation for the origin of the “Misc” dataset on the Unix station. The digitising of the blue line network had been contracted out and some data was deemed miscellaneous and put into a separate coverage. This provided an explanation for a considerable number of river mouths at inland locations, especially in Scotland. It was therefore necessary to collect the "Misc" data and merge it with the river network. Nodes had to be generated and renumbered. This was a very time-consuming task, taking well over one week to complete.

2.2.10 Errors found and corrected during editing of POINTS

The river network includes two sources of point data. First, the altitudes along arcs, as supplied by CEH Wallingford and second, the points representing nodes, sources and mouths which were generated by Mr D Hornby of CEH Dorset. Minor editing of the nodes/sources/ mouths has occurred when arcs have been removed or snapped to other arcs. The latter involves snapping the point to a new location and/or editing its identification number. When the network for a given Hydrometric Area has been fully processed and checked, nodes are renumbered from 1 to *n*, in order to reduce the file size for the attribute tables. This is done to optimise access to the attribute data as there is less to load into memory.

Major editing was necessary for the altitude data. First, null altitude values (-9999) were removed from each Hydrometric Area data. A script was then written to remove duplicate points. It is assumed that the macros used to get the altitude points along the river dealt with each arc in turn and, as a consequence, node points, other than source and mouth nodes, were duplicated because a node is shared by two or more arcs (or more than two arcs if a point represents a confluence). The removal of duplicate points greatly reduced the file size for each Hydrometric area and removed potential problems when querying the altitude data layer.



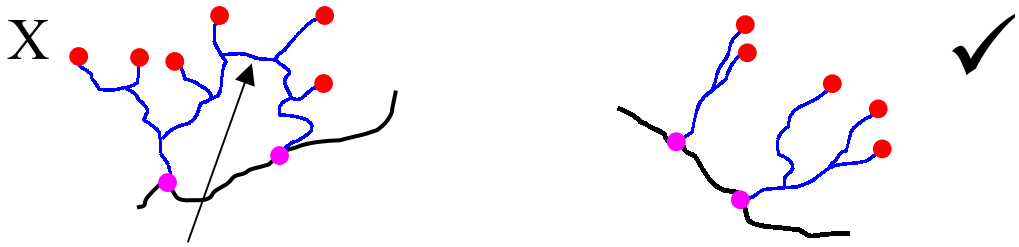
If arcs were removed or had vertices moved/removed, then any underlying height points became redundant, and these were also removed to reduce file size. This was achieved by selecting all height points that intersected arcs and then 'flipping' the selection such that heights which were not previously selected, became the selection. These were then removed.

Not all sources had their respective height points. Scripts were therefore created to output XY co-ordinates for all sources as a simple text file. These were e-mailed to Mr D Morris at CEH Wallingford, who processed them and returned a text file with each

XY co-ordinate plus an altitude value. Another script was then created to read the text file, convert the co-ordinates into a point shape, use this to check whether a height point was present and if not, to add the appropriate altitude value to the height coverage. This process took several weeks to complete.

2.2.11 A further check before the processing

To ensure that each network was processed correctly, it was sometimes necessary to make deliberate breaks. The processing scripts written by Mr D Hornby were designed to build routes between river mouths and their sources. If two catchments drain into separate mouths, a join between them at any part within the network (other than the mouth) can lead to sources in catchment A being treated as sources for catchment B and vice versa. (Note, however, that breaks were rejoined after such processing and any missed nodes were corrected).

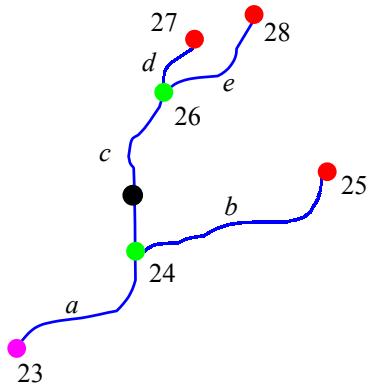


Network broken here and rejoined after processing.

2.3 Processing the GIS Network and Correcting Additional Errors

2.3.1 Processing the network

At the outset, it is important to reiterate why it is useful to 'process' the network. Essentially, this is because processing provides the basis for the efficient retrieval of map variable information for any given site in the future. When using ArcView to query the network for the map variables, the user enters the chosen site by clicking on the stream point. This point is then used to select the appropriate arc. Next, the arc attribute table is examined and the Fnode ID is returned. This ID is used to search the node attribute table and from this the ID of the nodes representing the source and mouth are returned. Using these ID numbers, the source and mouth points are returned and these are used with the ArcView network function *Findpath* to generate routes from the chosen site to the source and to the mouth. These routes define the main course of the river passing through the selected site and this is critical when calculating the slope and altitude of the site.



Node Attribute Table

Node ID	SourceID	MouthID
26	28	23
27	27	23
28	28	23

The chosen site (●) is on arc *c* with FNode 26. In the node attribute table above, node 26 has Source ID 28 and Mouth ID 23. These IDs are used to return the points, which ArcView uses to generate routes. The points in the source and mouth layers have the same ID number as their node in the node layer.

Before the network can be used to generate routes from the chosen site to the mouth and to the source, the actual processing of the network takes place in two stages. First the sources are calculated and then the mouths are calculated.

Calculating source and mouth of each node on the network

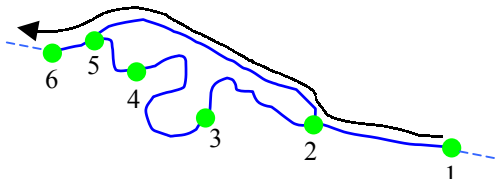
ArcView builds one list of points that represent all sources and a second for all mouths. Then, for each mouth in the list, ArcView attempts to generate a route from the mouth to all sources and successful routes are added to a third list. Given that, initially, the river network had been cut up into Hydrometric Areas, and each Hydrometric Area has more than one river mouth, not all mouths make routes to every source in the Hydrometric Area under consideration. The list of routes generated is then sorted by length of route. Starting with the shortest route, the route is used to select all the nodes along that route and these nodes are labelled with the ID number of the source node which the route points to. Nodes that share a common path are then overwritten with other source ID numbers with increasing route length. The process is then repeated using routes of increasing length until all mouths in the 'mouth list' have been processed; if a longer route passes through a node that has already been given a source ID number, the number is given the source ID number of the longer route. On completion each node is labelled with the source ID number of the longest mouth-to-source route passing through that node.

The mouth ID numbers of the nodes representing the mouth of each node on the river network are derived in the same way. The ID of the nodes representing the source and mouth are critical as it allows Arcview to define the main watercourse passing through a particular point (i.e. site) from which it will calculate the values of selected map-based variables for the site.

2.3.2 Errors generated during processing

Two types of error may be generated during the processing phase, one of which is easy to spot and correct, and a second which is more problematic

Routes are generated using the *FindPath* command, which works out the shortest path between two or more points. Clearly a straight side channel is shorter than a meandering main river course, so ArcView will generate routes along the side channel. In braided rivers, channels that are made up of several arcs can be missed out. As source/mouth node IDs were never assigned to the nodes which link these missed arcs, a simple query of the attribute table for all nodes which have null values will select the nodes that still need to be assigned a source/mouth ID in order to document the meandering route through nodes 1-6. In the example below, all the nodes 1-6 should be assigned the same source. In the automated process of determining the source and mouth of all nodes, nodes 1, 2, 5 and 6 will all have been given the same source and mouth IDs, but nodes 3 and 4 in the meandering section will have been missed. After the automatic processing, these missed nodes must then be assigned source and mouth IDs manually.

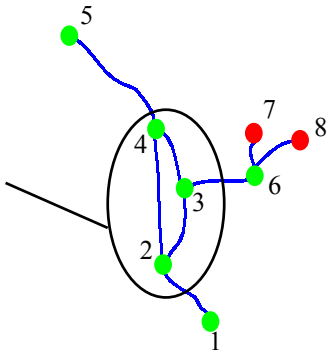


Initially, nodes 1, 2, 5 & 6 are labelled while 3 & 4 are missed

= Route taken by ArcView

One assumption is made during processing: the network is a simple dendritic pattern. Large parts of the network form grid-like networks, circular patterns or are highly braided. ArcView calculates the shortest path through all of these networks. In certain situations the network of nodes will point to incorrect sources, as illustrated in the example below. Hence it was necessary for Mr D Hornby to examine the entire network to check for specific patterns and ensure that they point to the correct source.

This braid is an island in the main channel, so nodes 1,2,3,4 & 5 should all point to the same source.



Nodes 1,2,3, 6 & 8 form a short route to source 8. The next longest route will be 1,2,4 & 5 to another source, but node 3 should also point to this source. Hence it needs adjusting.

2.3.3 Points to bear in mind when using the network

The user should be aware of the following situations:

Unable to select sites in BATCH mode

From the outset, the ArcView project was designed to enable a user to specify a list file of sites specified by their National Grid references and let the computer process them automatically, identifying their source and mouth and deriving values for select map-based variables. However, National Grid references are usually only specified to the nearest 100 m Easting and Northing, which is not always accurate enough for ArcView

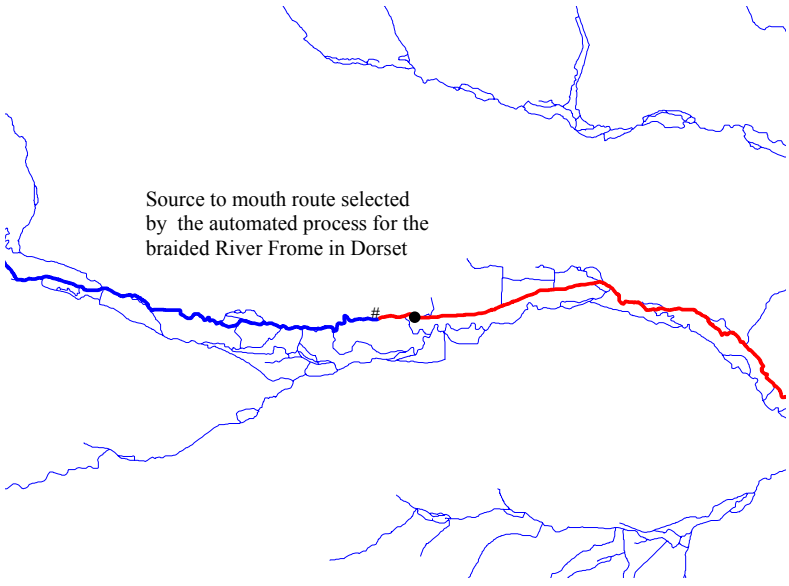
to select an arc on the GIS river network to initiate an upstream search. Early in the project, ArcView ignored a third of sites specified by their grid reference to 100 m accuracy because it could not select a single arc to start the search. Sites near a confluence (and hence two or more arcs) or not near any arc were the causes of these problems. However, improvements to the code now allow the user to choose to enforce the furthest upstream route when a site hits a confluence.

Selecting the wrong river in BATCH mode

Of even greater concern is the possibility of selecting the wrong arc at the beginning of the search. An inaccurate National Grid reference may, nevertheless, lie on a river (albeit the wrong one) and, as a consequence, ArcView starts a search and traces up the wrong tributary. These site selection problems can be avoided by using ArcView in an interactive mode where the user selects the exact location and can observe the route-finding process on the screen as a double-check.

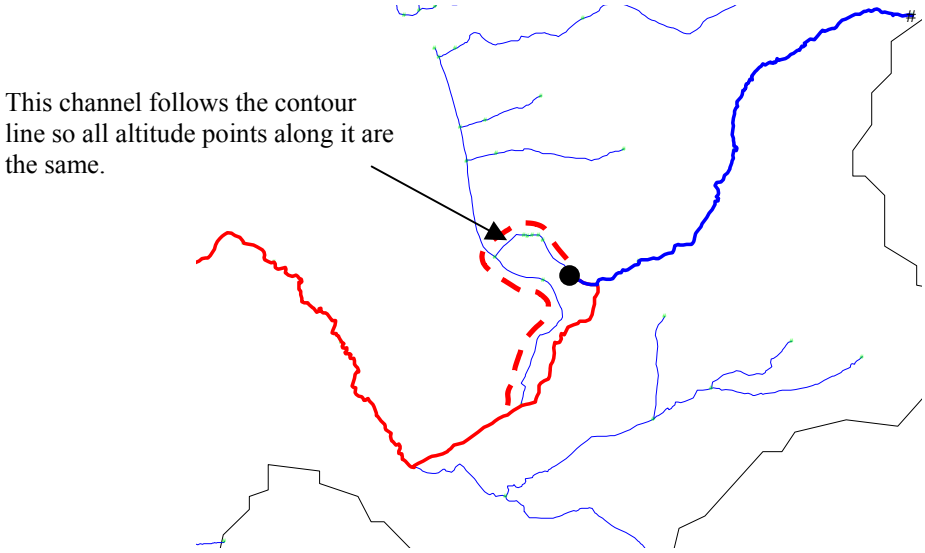
Tracing along the 'wrong' channel

At present, the dataset is not coded by size of channel so no priority can be given to following the ‘main’ channel during the route finding process. As previously noted, the network function *FindPath* calculates the shortest path between two or more points. Therefore, in braided channels, ArcView has a tendency to find routes which are not the main (more meandering) river course. This may have a minor consequence on the “distance to source” calculation. The example below shows the route chosen by ArcView on the River Frome, Dorset. It 'weaves' on and off the main channel as it runs to the source/mouth.



Routes that double back

In certain situations it is shorter for ArcView to go upstream and then back downstream to search for the mouth. This tends to happen in networks where channels "short circuit" dendritic patterns. The distance missed out by double backing routes tends to be quite small and affects only the "distance to mouth" calculation. In the figure below, the site in the centre (•) has source in the top right and the mouth is off to the left. Logic dictates that the route from the site to the mouth should have followed the dashed route instead of doubling back on itself.



Calculation of altitude at a site

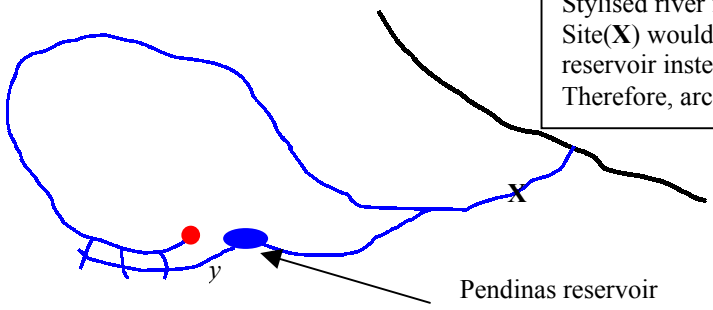
The altitude of a site is calculated by taking the average of the nearest upstream and downstream height points. If these points are far away from the site itself, then the averaged altitude for the site may be quite different from the altitude given on an OS map. Given the general patchiness of height points, especially in drainage networks, and the limited addition of height points in areas where arcs had to be merged into the data, poor average height data can be returned for some locations. Sites very near to lakes/lochs can give poorly-averaged heights due to no height points being on arcs that define the centre of lakes.

Slope calculation

ArcView attempts to calculate the slope of a site over the 1km stretch that extends 500 m upstream and 500 m downstream from the site. If the site is within 500 m of its source (or mouth) then the slope is calculated over the shorter distance plus the 500 m in the other direction.

Removal of arcs which 'short circuit' the genuine network

In a few cases, an arc has been removed from the network because it "short circuits" the rivers network. These arcs are channels that follow contour lines and effectively connect separate sources. In the example below from Hydrometric Area 67, it would be shorter for ArcView to trace up the *incorrect* tributary via the reservoir to the correct source at ●.

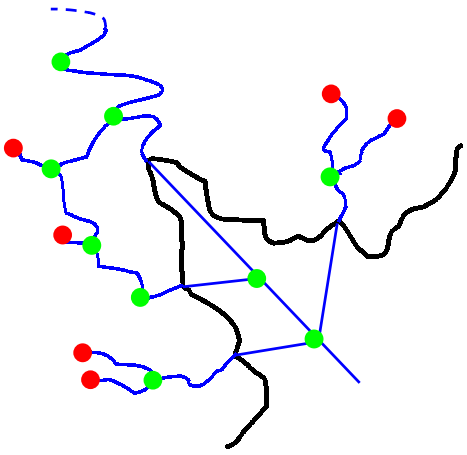


Stylised river from Hydrometric Area 67. Site(**X**) would trace to source via the reservoir instead of following main course. Therefore, arc (*y*) was removed.

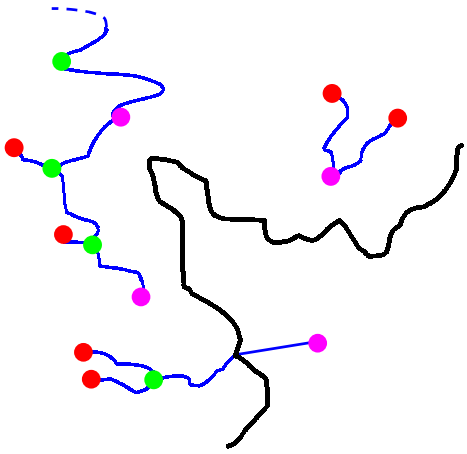
Downstream arcs

As previously stated, the entire network was cut into separate hydrometric areas using the "have centre within" criterion for each arc. This avoided the problem of having upstream "floating arcs" but had consequences for arcs at the downstream limit of the catchment. Many rivers' arcs were extended into the estuary and these extended arcs tended to have their centre outside the hydrometric boundary. Hence they were not included in the network. A consequence of this was that the river stops short of the hydrometric boundary.

Before cutting up into hydrometric areas



After cutting up

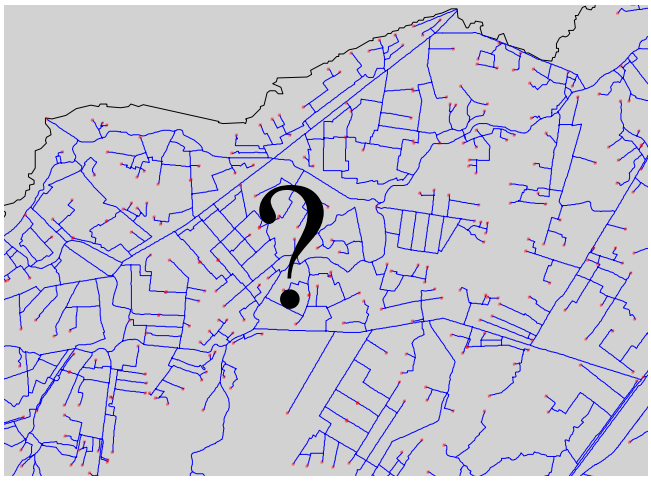


River networks which pose major problems for ArcView

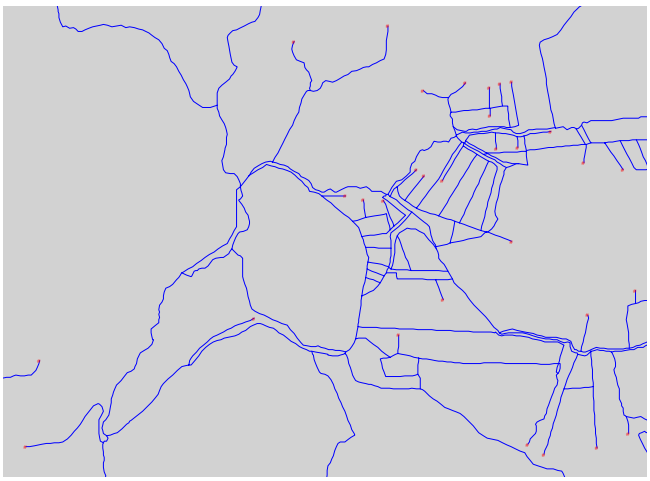
A relatively small number of Hydrometric Areas have grid-like drainage networks within them which pose major problems for ArcView and are probably not solvable. It may prove necessary to ignore sites that lie on such streams. Two examples are given below.

(A) Grid-like network from East Anglia (Hydrometric Area 33) and (B) Channels near Telford that form a circular pattern (Hydrometric Area 54)

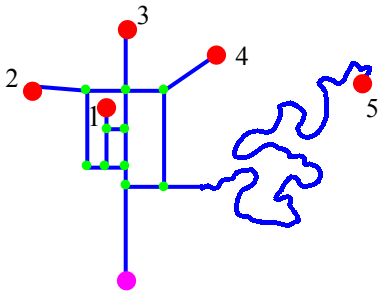
A



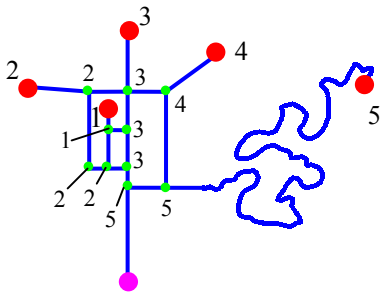
B



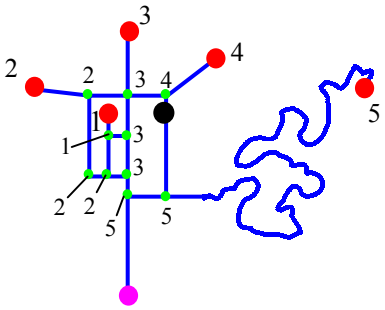
Why are grid-like networks not solvable? The scripts developed for assigning source/mouth IDs to nodes work well for normal dendritic patterns. Effectively, the nodes are "hard wired" to point to their sources as if you are travelling up the network starting from the mouth looking for a source. If an attempt is made to process a Hydrometric Area with a considerable amount of grid like network (e.g. Hydrometric Area 40), the river network is still processed and hard wired, but false results will be returned when you query the network for information on a specific site. The problem is that in low lying areas where there is little variation in height, a node which has been processed (hard wired) to point to a source from a mouth to source direction could equally point to another source when approached from another direction. An example, showing the network before processing. is given below. Note that all arcs point in a downstream direction.



During processing, routes are built from mouth to source and, in this case, five routes are built. In this example the length of route has been designed to relate to the source ID number, with 1 being the shortest route, and 5 the longest. Route 1 labels nodes to source 1 and this is overwritten by route 2, then route 3 and so on. The diagram below shows the green nodes and their assigned source IDs. In this example all nodes point to one mouth.

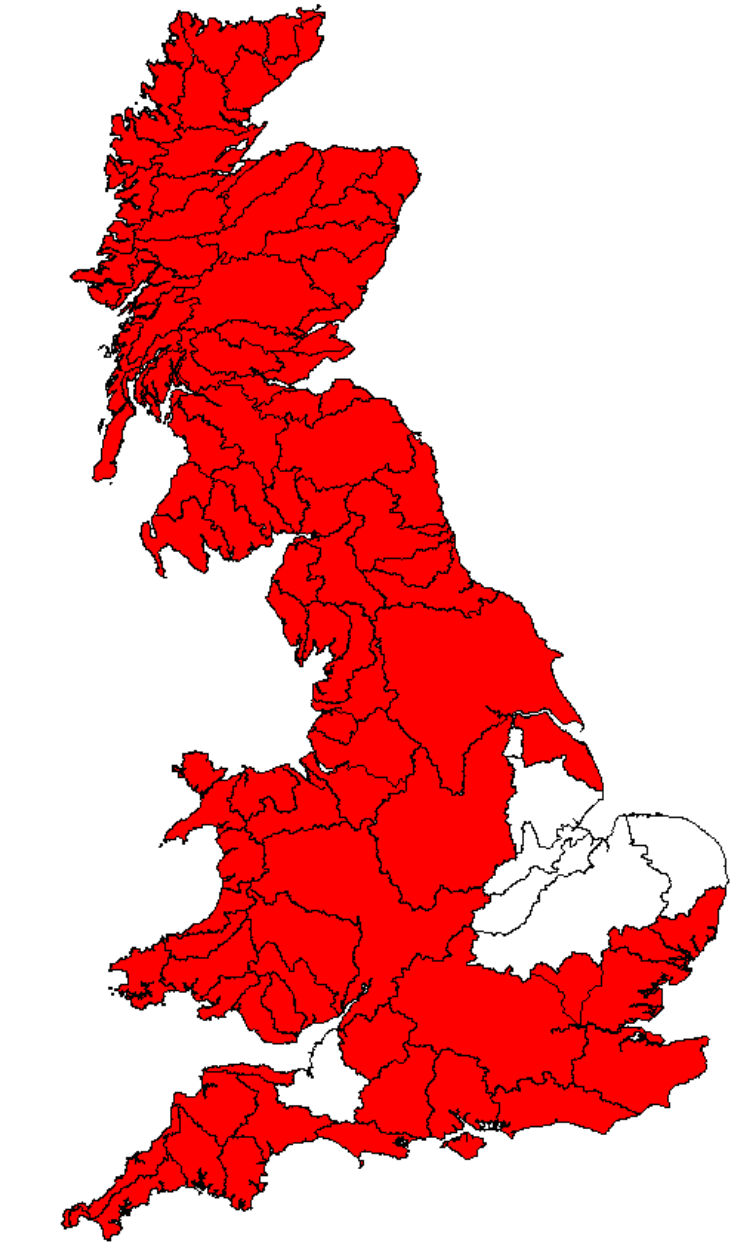


In the diagram below, if your chosen site (●) is as indicated, then the u/s node gives the source as 4, BUT if you could keep going upstream, then the furthest point away would actually be source 2.



The question remains: which source would have been traced using a map wheel?

Figure 2.1 shows the hydrometric areas of Great Britain whose blue line network has been corrected and which have been processed to determine the upstream source for all nodes in the area. The few hydrometric areas which have not been completed involve the low-lying drainage areas in East Anglia and Somerset. For some drainage systems in these very flat areas it is very difficult to determine the upstream source or even the direction of flow from just the blue line network and the available spot altitudes.



Hydrometric Areas which are only partially processed

Hydrometric Areas which have been edited and checked for errors

Figure 2.1: Map of the Hydrometric Areas (HA) of Great Britain showing which HAs have had their blue line network edited and quality controlled for errors (shaded) and the remaining eight low-lying HAs (unshaded) where problems remain due to grid-like drainage networks

3. COMPARISON OF GIS AND RIVPACS ESTIMATES OF CURRENT RIVPACS PREDICTOR VARIABLES

3.1 Acquisition of RIVPACS Predictor Variables from a GIS

Table 3.1 lists the environmental variables used in RIVPACS to predict the expected macroinvertebrate fauna at a site which can be acquired using the GIS described in Section 2. These variables can be generated within the GIS either manually or in batch mode if the sites of interest are supplied as a simple text file listing the NGRs. Preliminary testing of a sub-sample of sites demonstrated that batch processing can be unreliable. River sites can be missed or can generate inaccurate results because of incorrect selection of a river network site as a result of the inherent limitations of the 100 m accuracy of the NGRs. In view of this potential problem, the RIVPACS environmental variables for each of the RIVPACS III+ reference sites were obtained by manually entering their NGR. This causes ArcView to zoom into that location and the correct watercourse can then be selected manually by clicking the correct location on the computer screen.

Table 3.1: Methods of estimating current RIVPACS predictor variables from the GIS

RIVPACS variable	Acquired using automated GIS	How the GIS acquires the variable
National Grid Reference (NGR)	✓	The GIS returns the point as an NGR chosen to represent the site on the river. Note that this can differ from the NGR initially supplied.
Altitude of site	✓	The GIS returns the altitude of the site derived from averaging the nearest upstream and downstream altitude points along the river.
Distance from source	✓	The GIS calculates the furthest point upstream from the site, *using the shortest route*. In a simple dendritic system (e.g. River Ribble) the shortest route is the major channel. In a highly managed system (e.g. River Thames) the shortest route will be along straight side channels instead of the meandering major channel.
Slope at site	✓	The GIS attempts to calculate the slope at the site over a distance of 1 km. It searches upstream 500 m and downstream 500 m for an altitude point and then divides the altitude difference by the distance between the altitude points. In cases where the site is less than 500 m from the source or mouth then the altitude difference is divided by the lesser distance.
Discharge category	×	At present, the GIS cannot automatically provide the discharge category. This is because the discharge layer of information in the GIS was digitised at a different spatial scale, accuracy and level of detail and consequently does not overlie the blueline network. However, discharge can be acquired by loading the discharge category coverage layer into the GIS on top of the corrected blue line river network and then manually selecting the appropriate river and hence discharge on the discharge layer.

At the time of this analysis, it had not been possible using the available GIS routines to identify the upstream route and hence source for sites in some very flat parts of England. These problematic areas included 49 of the 614 reference sites, which meant that it was not possible to use the automated GIS routines to calculate their altitude, distance from source or slope at site. Therefore, the remaining 565 sites are used in the comparisons of GIS and RIVPACS estimates of these variables.

3.2 Altitude of Site: Comparison of Estimates

Altitude at a site was measured with the GIS using the method given in Table 3.1. Altitude is recorded in metres, with a minimum value of 1 m.

Altitude of site measured by GIS was greater than the original RIVPACS values for 40% of sites and less for 59% of values, suggesting some slight overall tendency for GIS values to be less than the manually estimated original altitude values (Table 3.2). However, less than 2% of sites had estimates differing by more than 20 metres.

Table 3.2: Percentage of reference sites within each size class of difference in estimate of altitude of site; difference = GIS value minus current RIVPACS value

Difference in altitude (m)	% of sites	cumulative % of sites	% difference	% of sites	cumulative % of sites
<-20	0.9	0.9			
-20 : -10	3.7	4.6	<-50	1.2	1.2
-10 : -5	10.6	15.2	-50 : -20	7.1	8.3
-5 : -2	16.6	31.8	-20 : -10	6.0	14.3
-2 : -1	11.0	42.8	-10 : -5	10.5	24.8
-1 : 0	17.1	59.9	-5 : 0	35.2	60.0
0 : 1	17.2	77.1	0 : 5	25.8	85.8
1 : 2	6.4	83.5	5 : 10	4.1	89.9
2 : 5	9.7	93.2	10 : 20	4.3	94.4
5 : 10	3.4	96.6	20 : 50	2.5	96.7
10 : 20	2.5	99.1	>50	3.3	100.0
>20	0.9	100.0			

Within the RIVPACS III+ software, estimation of the expected fauna at a site is based on the use of multivariate equations, referred to as discriminant functions derived from a statistical multiple discriminant analysis (MDA) predicting the biological group of the reference sites from their values for key environmental variables (Clarke *et al.* 1996). In these discriminant functions, several variables are used in the equations in their logarithmic form, rather than their absolute values. These variables are altitude at site, distance from source, slope at site, water width and water depth. There are two reasons for this: firstly, using these variables in their logarithmic form gives better discrimination between the biological site groups and hence improved predictions of the expected fauna, and secondly, it makes ecological sense to assume that unit increase in any of these variables has less effect on the expected fauna as the value of the variable increases. For example, Furse (2000) highlights that numerous taxa are restricted to headwater streams within one or two kilometres of source and that the

macroinvertebrate composition can change rapidly over the first few kilometres of a stream. However, the same type of fauna would, on average, be expected in streams 50 km and 60 km from source. Similarly, changes in fauna are expected to be greater for increases in stream slope from 1 to 5 metres per km (m km^{-1}) than for increases from 30 to 35 m km^{-1}, or for increases in water depth from 10 to 100 cm compared to increases from 5.1 to 6 m.

The consequence, in the current context, of using the logarithmic form of these environmental variables in the predictive equations, is that a particular proportional change, rather than absolute change, in the estimates for any these variables is more likely to have a similar effect on changes in the expected fauna. Thus, it is best to analyse differences between the GIS-based values and the original RIVPACS values on a logarithmic scale and highlight differences above particular percentage changes from the original value. All logarithms are to base 10 (log_{10}).

Figure 3.1(b) therefore plots the GIS value for log_{10} altitude versus the RIVPACS log_{10} value for the RIVPACS reference sites and highlights the 14% of sites for which the difference is greater than 20% of the original RIVPACS value (Table 3.2).

All the sites for which the discrepancy is greater than 20% had RIVPACS values for altitude of less than 200 m. Moreover, all the 26 sites (4.5%) for which the discrepancy was >50% had RIVPACS altitude values of no more than 25 m. and the largest percentage differences were all for sites with RIVPACS altitude values of 15 m or less (Figure 3.1). Where the difference in altitude values was greater than 90% the map was re-examined and, if necessary, the GIS was investigated; the reasons for each difference are given in Table 3.3. There were seven sites with reasonable RIVPACS altitude values of <5 m whose GIS values were 2-3 times higher but still ≤8 m. There were just four sites whose RIVPACS altitude value was confidently judged to be wrong (Table 3.3).

Table 3.3: Investigation of sites with very large percentage differences (% diff.) between the RIVPACS value (R) and the GIS-based estimate of altitude of site

Site code	River name	Site name	R	GIS	% diff	Explanation
3413	Tees	Over Dinsdale	4	8	100	Reasonable RIVPACS estimate for very low altitude. GIS gives an altitude that is 2-3 times the original RIVPACS value.
4313	Carron	New Kelso	2	5	150	" "
5509	Stour/ Great Stour	Fordwich	1	3	200	" "
8517	Piddle	Wareham	2	5	150	" "
H107	Shiel	Shiel Bridge	2	5	150	" "
2409	Great Eau	Theddlethorpe -All-Saints	2	5	150	Very flat area. RIVPACS estimate of altitude would have to be guessed.
5309	Lymington	Boldre Bridge	2	5	150	RIVPACS has taken altitude from bridge (2 m), GIS has calculated an altitude higher than this, so they are both wrong, assuming OS is correct.
0713	Avon	Staverton Weir	15	29	93	RIVPACS value incorrect, GIS closer to the true height
H105	Unnamed	Mon	15	48	220	" "
5009	Otter	Newton Poppleford	12	23	92	" "
1413	Lee	Enfield Weir	7	17	143	RIVPACS possibly misidentified altitude because of lack of contours and other cartographic features obscuring river and contours.
1611	Teifi	Llechryd	7	15	114	GIS has incorrect height values labelled along the Teifi.

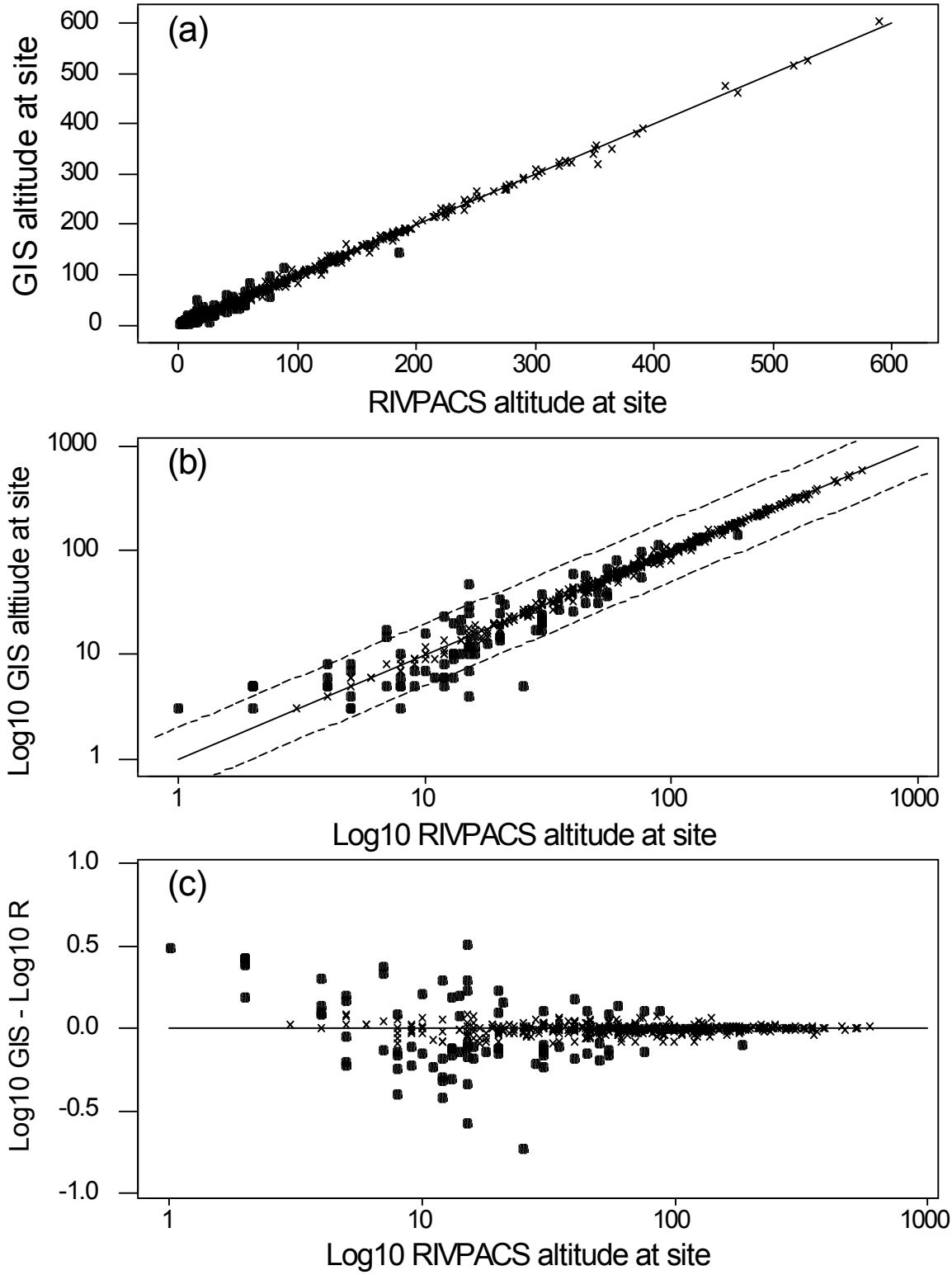


Figure 3.1: Comparison of altitude at site obtained from GIS with the current manually-obtained value (R) in RIVPACS for the RIVPACS III+ reference sites on (a) untransformed and (b)-(c) log_{10} transformed scales. Sites for which the GIS value differed from the RIVPACS value by >20% are denoted by •; dotted lines in (b) indicate GIS value was either double or half the RIVPACS value

3.3 Slope at Site: Comparison of Estimates

Slope at site is recorded in metres per kilometre (m km^{-1}) to one decimal place. The slope at site was measured within the GIS using the method given in Table 3.1 and is thus a measure of the average slope over the river stretch extending from 500 m upstream to 500 m downstream. This differs from the method prescribed for RIVPACS in the procedures manual of Murray-Bligh *et al.* (1997), which states that slope should be estimated from the difference in height between the nearest upstream and downstream altitude contour lines on the 1:50,000 OS map, divided by the stream distance between the contours. It was not possible to use the GIS to calculate slope using the same method as done manually because CEH does not currently have the OS altitude contour maps available within its national GIS system. CEH has detailed altitudes at frequent points along the blue-line river network, as derived and supplied by CEH Wallingford. Given this difference in definition and distance over which the slope is usually measured, the very large differences between the two methods in estimates of slope which occurred for some reference sites were to be expected (Figure 3.2).

There is no major tendency for the GIS estimate of the slope at a site to be higher than the RIVPACS value; 57% of sites had a lower value of slope when estimated by the GIS method (Table 3.4). However, for 16 sites (2.8%) the GIS slope was less than 10% of the original RIVPACS value, whilst for 46 sites (8.2%) the GIS slope was more than double the RIVPACS value.

Table 3.4: Percentage of reference sites within each size class of difference in estimate of slope at site (m km^{-1}); difference = GIS value minus current RIVPACS value

Difference in value of slope at site (m km^{-1})	% of sites	cumulative % of sites	% difference	% of sites	cumulative % of sites
<-20	2.1	2.1	<-90	2.8	2.8
-20 : -10	2.0	4.1	-90 :-50	8.0	10.8
-10 : -5	3.5	7.6	-50 :-20	20.4	31.2
-5 : -2	6.9	14.5	-20 :-10	9.6	40.8
-2 : -1	7.3	21.8	-10 : -5	6.4	47.2
-1 : 0	35.1	56.9	-5 : 0	9.7	56.9
0 : 1	23.4	80.3	0 : 5	3.2	60.1
1 : 2	6.6	86.9	5 : 10	6.0	66.1
2 : 5	6.5	93.4	10 : 20	8.5	74.6
5 : 10	3.2	96.6	20 : 50	9.9	84.5
10 : 20	2.0	98.6	50 :100	7.3	91.8
>20	1.4	100.0	>100	8.2	100.0

The sites for which the GIS value for slope was several times (>6) greater than the RIVPACS value were investigated (Table 3.5). The RIVPACS value for slope should be checked at those sites where there was major unexplainable difference in slope estimates.

If the OS altitude contour GIS layer was available, it would be relatively easy to add software routines to the GIS to calculate an estimate of the slope at any site using the same method as currently done manually for RIVPACS. Obviously, the best distance

over which to calculate slope at site and the best method of calculating it is the one which helps give the best predictions of the expected macroinvertebrate fauna.

Table 3.5: Investigation of sites with very large percentage differences (% diff.) between the RIVPACS value (R) and the GIS-based estimate of slope at site (m km^{-1})

Site code	River name	Site name	R	GIS	% diff	Explanation
3403	Tees	Cauldron Snout	3.3	119.3	3515	Site is immediately d/s of Cow Green Reservoir. RIVPACS possibly misidentified slope because of other cartographic features obscuring contours and river. Also GIS calculated slope over a distance of 0.48km instead of 1km as there is a poor coverage of height points along the arcs and none on centre line of reservoir.
4303	Lair	Achnashellach Lodge	5.7	71.9	1161	Site is immediately u/s of Loch Dùghaill. RIVPACS possibly misidentified slope because of other cartographic features obscuring densely packed contours. Also GIS calculated slope over a distance of 0.7km instead of 1km as are no height points along the arcs on the centre line of Loch.
5805	Eastern Cleddau	Llawhaden	0.7	6.9	886	The GIS method of calculating slope gives a greater slope value.
9481	Walkham	Merrivale	2.5	27.3	992	" "
H105	Un-named	Mon	9.1	195.6	2049	RIVPACS possibly misidentified slope because of densely packed contours. Also RIVPACS distance to source incorrect. GIS u/s limit is 210m while d/s is 10m, hence greater slope value.
H106	Un-named	Craig Ghobhair	50	365.1	630	RIVPACS possibly misidentified slope because densely packed contours. The GIS method of calculating slope gives a greater slope value.
TW03	Eden Water	A6089 Bridge	0.1	4.6	4500	Site is in a saddle. The GIS method of calculating slope gives a greater slope value.

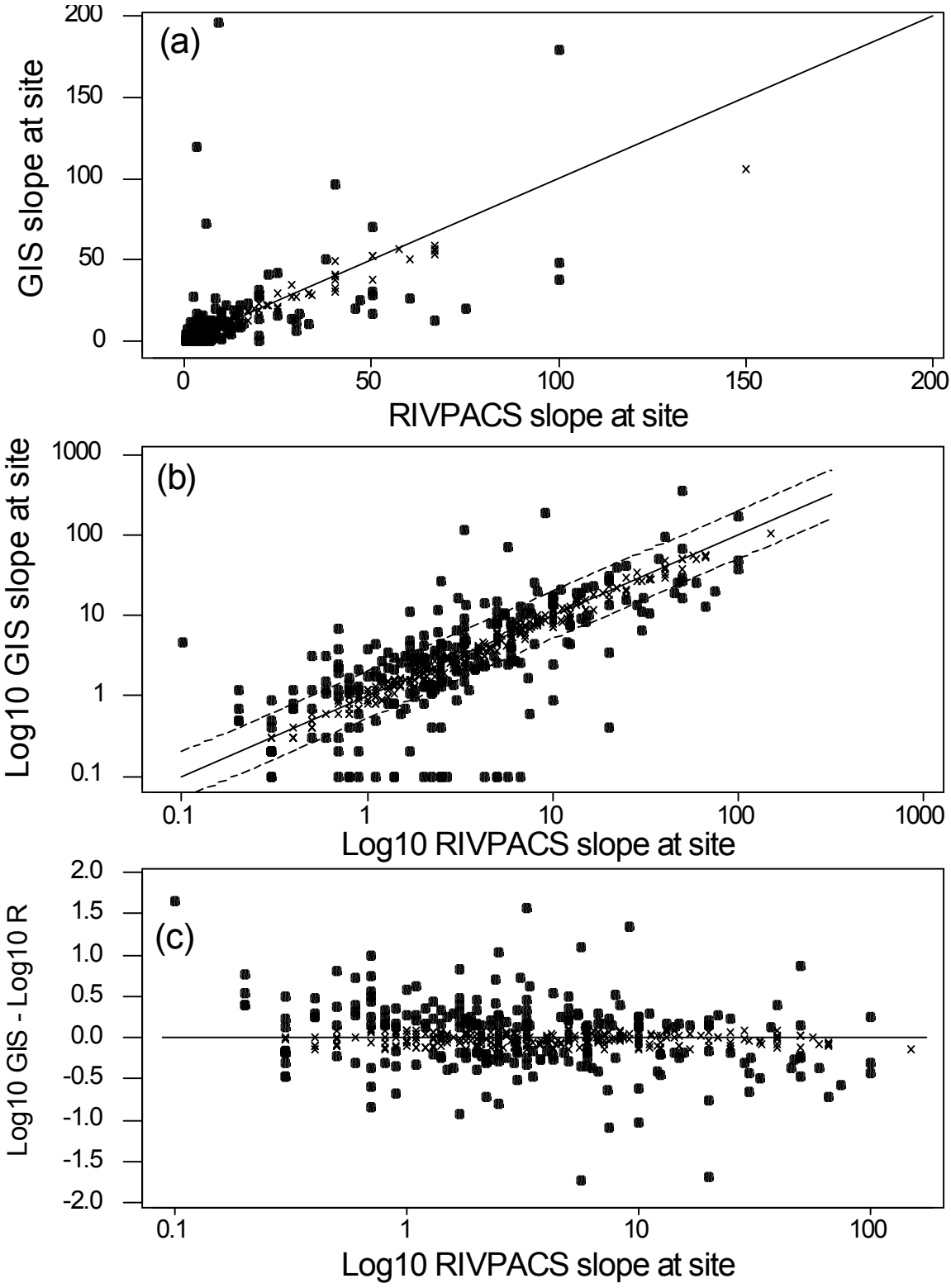


Figure 3.2: Comparison of slope at site obtained from GIS with the current manually-obtained value (R) in RIVPACS for the RIVPACS III+ reference sites on (a) untransformed and (b)-(c) $\log_{10}$ transformed scales. Sites for which the GIS value differed from the RIVPACS value by >30% are denoted by •; dotted lines in (b) indicate GIS value was either double or half the RIVPACS value. One extreme outlier site had a slope of 50 m km^{-1} for RIVPACS and 365 m km^{-1} by GIS

3.4 Distance from Source: Comparison of Estimates

Distance from source was measured within the GIS using the method given in Table 3.1; it is recorded in km to the nearest 100 m.

Distance from source measured by GIS was greater than the original RIVPACS values for all except 10% of the reference sites (Table 3.6, Figure 3.3). For 10% of sites the GIS value was more than 50% greater and for 3% of sites it was more than double the RIVPACS distance from source value (Table 3.6).

In retrospect, this is not unexpected, as the original measurements of distance from source for the RIVPACS reference sites were done manually using a measuring wheel on 1:50000 scale OS maps. This method is likely to ignore or under-estimate the length of some minor loops or meanders and treat the river course as straighter than it actually is. Hence, the manual method is likely to underestimate the true blue line distance from source. The RIVPACS procedures manual (Murray-Bligh *et al.* 1997) states that distance from source should be measured as "the distance along the watercourse (in kilometres, to the nearest 0.1 km) between the site and its furthest source, regardless of whether that source is on a tributary known by a different name". The GIS method will always measure distance to the furthest upstream source. However, with the original manual method, the recorder may have incorrectly chosen a different upstream source which is a shorter river distance from the site.

Table 3.6: Percentage of reference sites within each size class of difference in estimate of distance from source; difference = GIS value minus current RIVPACS value

Difference in value of distance from source (km)	% of sites	cumulative % of sites	% difference	% of sites	cumulative % of sites
<-10	0.7	0.7			
-10 : -5	0.4	1.1	<-50	1.4	1.4
-5 : -2	0.7	1.8	-50 : -20	1.4	2.8
-2 : -1	1.0	2.8	-20 : -10	0.9	3.7
-1 : 0	7.3	10.1	-10 : -5	1.3	5.0
0 : 1	22.1	32.2	-5 : 0	5.1	10.1
1 : 2	11.7	43.9	0 : 5	9.7	19.8
2 : 5	24.4	68.3	5 : 10	22.1	41.9
5 : 10	18.1	86.4	10 : 20	20.2	62.1
10 : 15	6.5	92.9	20 : 50	27.5	89.6
15 : 20	3.7	96.6	50 : 100	7.2	96.9
>20	3.4	100.0	>100	3.2	100.0

An investigation of several sites for which the GIS value was more than double the original RIVPACS value for distance from source revealed that in all cases the original RIVPACS value was wrong because it was based on the wrong upstream source and did not measure distance to the furthermost upstream source (Table 3.7).

Table 3.7: Investigation of sites with very large percentage differences (% diff.) between the RIVPACS value (R) and the GIS-based estimate of distance from source. (In all cases the RIVPACS value is wrong because it did not measure distance to the furthest source.)

Site code	River name	Site name	R	GIS	% diff
501	Avill	Wheddon Cross	1	3.8	280
1081	Hammer's Pond Tributary	Carter's Lodge	0.3	2.5	733
1901	Perry	Perry Farm	4	13	225
3601	Wansbeck	Kirkwhelpington	5	11.5	130
4403	Loanan	D/S Loch Awe	1	5	400
4701	Halladale	Forsinard Lodge	4	11.6	190
4805	Thurso	Westerdale	16	43.4	171
4885	Unnamed	Westerdale	0.6	4.4	633
5401	Beult	Hadman's Place	8	19.5	144
5844	Unnamed	Dunhampton Farm	0.4	3.1	675
6501	Mounton Brook	Bully Hole Bottom	0.8	1.8	125
6848	Unnamed	Woolland	0.3	0.8	167
WE05	Morlas Brook	D/S Glyn Morlas	2	11.3	465

In summary, it would be sensible to use the GIS routine in the future in order to measure distance from source for any site. The GIS-based values should also be used for the RIVPACS reference sites themselves, but this would entail developing new, albeit probably only slightly different, discriminant functions for use within the RIVPACS software, from which to predict the expected fauna at any site. In section 4, the effect of using the new GIS values for the RIVPACS discrimination and predictions is assessed.

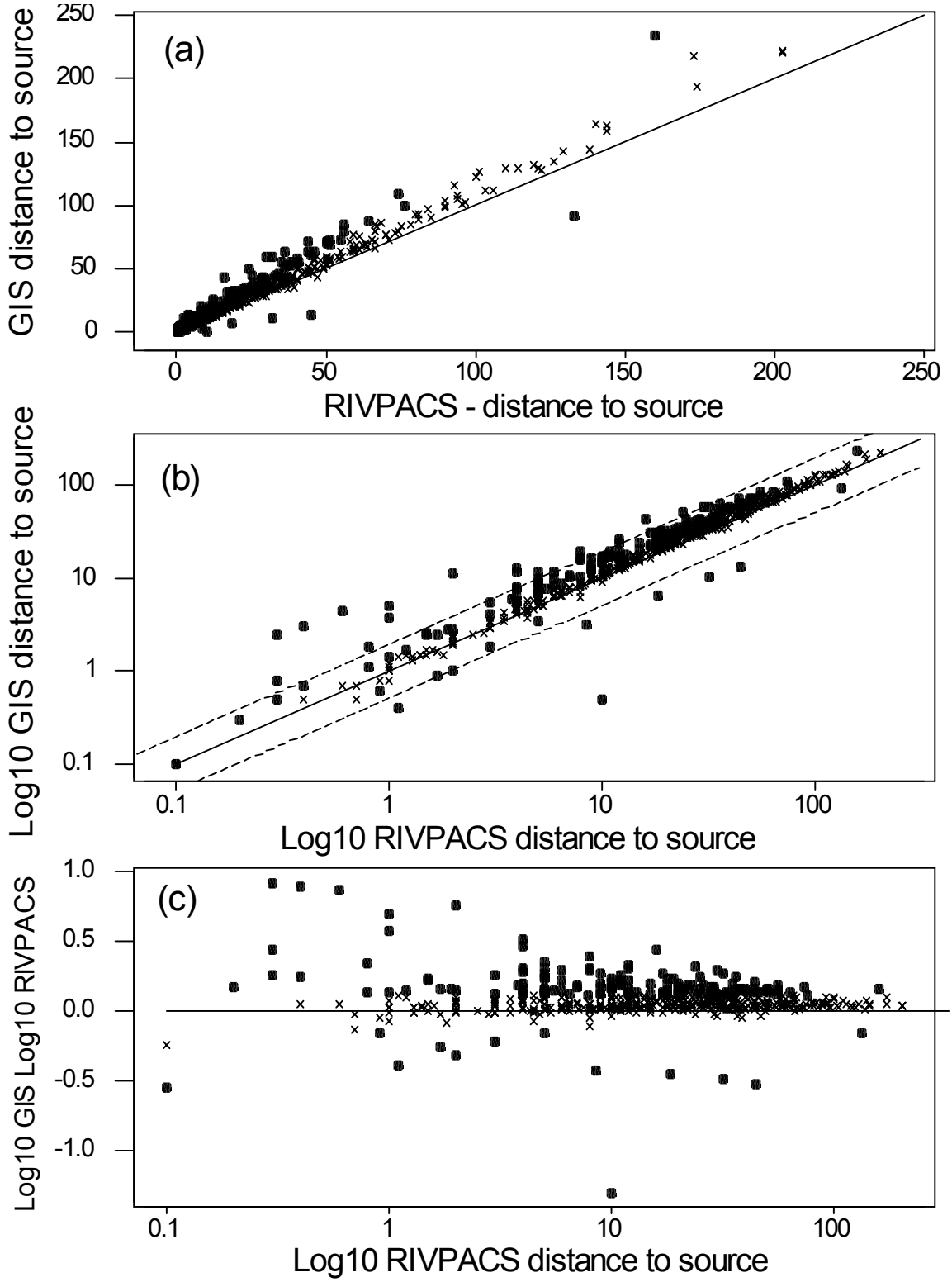


Figure 3.3: Comparison of distance from source obtained from GIS with the current manually-obtained value (R) in RIVPACS for the RIVPACS III+ reference sites on (a) untransformed and (b)-(c) $\log_{10}$ transformed scales. Sites for which the GIS value differed from the RIVPACS value by >30% are denoted by •; dotted lines in (b) indicate GIS value was either double or half the RIVPACS value

3.5 Discharge Category: Comparison of Estimates

The estimated annual mean discharge for a site is included in RIVPACS as a predictor variable using logarithmic (i.e. doubling) discharge categories (1-10) (Table 3.8).

The methods prescribed for estimating the annual mean discharge for use in RIVPACS predictions are given in section 2.5.5 of Murray-Bligh *et al*. (1997) which states "The information *(i.e. mean annual discharge)* should be obtained by direct gauging or the Institute of Hydrology's Micro Low Flow system if available to you. If information about discharge category cannot be obtained, current velocity measured at the site may be used as a substitute. Micro Low Flows may give unrealistic estimates under certain conditions, so it is recommended that you seek the advice of a hydrologist if you use this system. The discharge categories marked on national river quality maps (e.g. National River Authority, 1994) are known to include inaccuracies."

Despite these reservations about the use of discharge categories on national maps, the digital version of the Environment Agency's discharge map was the best data available as a GIS network layer. This discharge layer was superimposed on the blue line network and the discharge category for each of the RIVPACS reference sites was obtained by selecting the most appropriate nearby river site and hence discharge category on the discharge layer, as detailed in Table 3.1. It is important to recognise that the blue line network and the discharge river network lines were originally digitised from different scale maps and hence do not perfectly overlie one another.

Table 3.8: Discharge categories in RIVPACS

Discharge category	Mean annual discharge (m^3s^{-1})
1	< 0.31
2	0.31 – 0.62
3	0.62 – 1.25
4	1.25 – 2.5
5	2.5 – 5.0
6	5 – 10
7	10 – 20
8	20 – 40
9	40 – 80
10	> 80

The discharge map was only available for England and Wales. This meant that it was not possible to obtain a GIS-based discharge category for any of the 189 RIVPACS reference sites in Scotland. Of the 425 reference sites in England and Wales, the GIS method did not obtain a discharge category for 49 sites in low-lying areas for which the upstream source and route was undetermined and there were a further three sites for which the discharge category was unobtainable. This left 373 sites for which the discharge category was estimated using the GIS.

Table 3.9 compares the GIS discharge category for each site with its original RIVPACS discharge category. Between half and three-quarters of the sites given each discharge category within RIVPACS were given the same category using the GIS approach. Overall 62% of sites were assigned to the same category by both methods (Table 3.10).

Moreover, a further 17% of sites were assigned to one category lower using GIS than in RIVPACS and 12% to one category higher. Thus 9% of sites were assigned to a discharge category using the GIS method which was more than one category different from their original RIVPACS discharge category. Based on a sensitivity analysis of the effect of errors in measuring the RIVPACS environmental variables on estimates of the expected values of BMWP indices and hence Ecological Quality Indices (EQIs), Clarke *et al* (1994) concluded that for very small streams with discharge category 1-2, it needed to be recorded correctly, whilst for all larger streams, an error of ± 1 category had little effect on assessments of site quality.

Table 3.9: RIVPACS reference sites in England and Wales cross-classified by their GIS and RIVPACS discharge category (total *n* = 373)

		GIS value for discharge category									total	% same
		1	2	3	4	5	6	7	8	9		
RIVPACS value for discharge category	1	45	13	7							65	69
	2	11	23	6	3						43	53
	3	2	11	23	4	3	1				44	52
	4		2	12	36	9	3				62	58
	5		1	2	11	31	5	1			51	61
	6			2	1	4	30	3			40	75
	7			1	2		11	24	4		42	57
	8					1		1	11	2	15	73
	9								3	8	11	73
	total	58	50	53	57	48	50	29	18	10	373	62

Table 3.10: Number of RIVPACS reference sites in England and Wales given a higher or lower discharge category using GIS than their original RIVPACS discharge category (total *n* = 373)

Difference in discharge category: GIS minus RIVPACS	*n* sites	% of sites	histogram of % of sites	% of sites with GIS category:
-4	1	0.3	*	
-3	6	1.6	*	lower = 20.9%
-2	7	1.9	*	
-1	64	17.2	*******	
0	231	61.9	**********	same = 61.9%
1	46	12.3	*****	
2	17	4.6	**	higher = 17.2%
3	1	0.3	*	

In summary, the GIS-based method of determining the discharge category of a site from the digitised version of a national map of discharge categories agrees within ± 1 category with the estimate based on manual reading of the printed form of the same discharge map for over 90% of sites. In such cases, the two methods will lead to similar estimates of the expected fauna in RIVPACS predictions. For the remaining sites, where the discrepancy is greater, the current manually-obtained discharge category should be re-checked.

4. EFFECT OF USING GIS-DERIVED VALUES OF CURRENT PREDICTOR VARIABLES IN THE MDA

4.1 Introduction

The current suite of RIVPACS environmental variables is the subset of variables, selected from a larger initial set, that gave the best ability to predict the biological group of the 438 RIVPACS II reference sites using the multivariate statistical technique of multiple discriminant analysis (MDA) (Moss *et al* 1987). In the development of RIVPACS III, the extended set of 614 reference sites were re-classified into 35 biological groups, but the same suite of environmental variables were re-used to derive new predictive discriminant function equations.

As stated in Section 2, at the time of this analysis, some low-lying hydrometric areas with very flat drainage sections could not be processed by the CEH Dorset GIS. Thus, GIS values for environmental variables were not available for any of the 49 RIVPACS reference sites within these hydrometric areas. To assess the effect of using the GIS values for the current RIVPACS environmental variables on RIVPACS predictions using the MDA, it was considered important to involve all the 614 RIVPACS reference sites. Therefore, in the case of the 49 sites with no GIS values for altitude at site, slope at site, distance from source and discharge category, the current RIVPACS values were used instead.

4.2 Results

All the current RIVPACS environmental variables are expected to have some ability to discriminate between the RIVPACS biological site groups because this is the purpose for which they were originally selected. The abilities of each of the variables, when used on their own, to discriminate between the 35 site groups were fairly similar (Table 4.1). Log alkalinity was marginally the best single variable.

Table 4.1: Ability of each environmental variable, when used on its own, to predict the TWINSPAN biological group of the 614 RIVPACS reference sites.

	% of sites classified to correct group using	
	Current RIVPACS values	GIS values
Log alkalinity	15.6	
Longitude	14.2	
Log distance from source	13.4	13.2
Alkalinity	13.5	
Mean substratum	13.0	
Log slope	13.0	14.3
Mean air temperature	12.7	
Discharge category	12.4	12.2
Air temperature range	12.2	
Latitude	12.1	
Log stream width	11.2	
Log stream depth	11.1	
Log altitude	10.1	10.0

When the GIS values for distance from source, altitude at site, slope at site and discharge category were used instead of the current RIVPACS values for these variables, the discriminatory ability of each variable was similar; although slope at site was slightly more effective using the GIS values. This is interesting because this is the only variable which was, *for convenience of estimation*, defined differently using GIS. The GIS value for slope at site is estimated as the average slope over the 500 m upstream and 500 m downstream stretch, rather than the slope between the nearest upstream and downstream altitude contours.

Table 4.2 shows the results of a stepwise multiple discrimination technique, using the SAS software procedure PROC STEPDISC (SAS, 1989). At each step, this technique adds to the predictor set the variable which gives the greatest statistically significant improvement in discriminatory power, as measured by a multivariate analysis of variance F test, after allowing for the effect of the variables already included. One practical measure of the discriminatory power of a set of variables is the percentage of sites which are allocated to the correct site group using the discriminant function equations based on these variables (Moss *et al* 1987, Clarke *et al*. 1996). Using this method of estimating discriminatory power, which is known as the re-substitution method, the percentage of sites assigned to the correct group tends to increase as extra variables are included. However, once all the effective variables have been included, adding further variables can give slight reductions in the percentage allocated to correct group, as happened in Table 4.2 at step 11 when adding the Latitude variable. In general, the re-substitution method tends to over-estimate the overall effectiveness of the discriminant functions when several variables are involved. A better estimate of true effectiveness is to carry out the discrimination using all the RIVPACS reference sites except one, test whether the derived discriminant functions can correctly allocate the omitted site to its correct group, and then repeat this omitting each site in turn. Using this approach, which is referred to as the cross-validation method, the estimate of the percent allocated to the correct site group reaches an asymptote when the unused variables add no extra discriminatory power. This is best assessed by comparing the columns headed "re-substitution" and "cross-validation" in Table 4.2.

Table 4.2: Stepwise discrimination showing the order of selection of environmental variables to predict the TWINSPAN biological group of the 614 RIVPACS reference sites using (a) current RIVPACS values and (b) GIS values for the variables marked *

Order of variable selection by stepwise multivariate ANOVA		(a) Current RIVPACS values		(b) GIS values	
		Cumulative %classified to correct group by:		Cumulative %classified to correct group by:	
		re-substitution	cross-validation	re-substitution	cross-validation
1	Log alkalinity	15.6	15.6	15.6	15.6
2	Log distance from source *	24.3	22.6	24.6	22.8
3	Mean substratum	30.0	28.3	31.1	28.3
4	Mean air temperature	37.5	33.9	38.8	34.4
5	Alkalinity	39.4	36.3	41.9	37.1
6	Discharge category *	41.2	37.0	41.4	34.5
7	Log stream depth	43.3	37.8	43.5	36.5
8	Longitude	46.3	39.6	45.4	37.6
9	Log altitude *	46.4	40.4	47.4	39.3
10	Log slope *	49.5	40.6	48.9	39.3
11	Latitude	49.2	41.4	51.5	41.0
12	Air temperature range	49.7	41.0	52.1	41.0
13	Log stream width	51.3	41.2	52.8	40.7

For example, after allowing for the effect of 'log alkalinity', the variable 'log distance from source' gave the greatest improvement, such that using these two variables only in the discriminant functions assigned 24.3% of the reference sites to their correct group; using the cross-validation method the percentage correctly assigned is estimated to be slightly lower at 22.6%. The difference between the discriminatory power estimates from the re-substitution and cross-validation methods increase as more variables are added and the re-substitution method starts to "over-fit" by making use of idiosyncrasies in the dataset. (This is the same type of over-fitting problem as that which occurs when using multiple regression with too many variables compared to the number of observations).

The best prediction of TWINSPAN groups, as assessed by cross-validation, occurred when all (or at least 11) of the 13 current RIVPACS III+ preferred option 1 variables were included in the discrimination. With all 13 variables included and using the current RIVPACS values for the 614 reference sites, the percentage of reference sites allocated to the correct TWINSPAN group is estimated to be 51.3% using the re-substitution method, but only 41.2% using the cross-validation method.

Using the GIS values for distance from source, discharge category, altitude and slope at site did not give any improvement in the percentage of sites allocated to the correct TWINSPAN group; the corresponding percentages correct were 52.8% and 40.7% using the re-substitution and cross-validation methods respectively.

4.3 Discussion

At each stage in the stepwise addition of extra variables, the percentage correctly allocated to group was similar regardless of whether the GIS or original RIVPACS values were used for these four variables (compare left and right hand sides of Table 4.2). No new subsets of these variables involving one or more variables with values derived by GIS gave any improvement in discrimination.

All these comparisons of the effectiveness of different combinations of variables have been assessed using the standard statistical methods used in all discrimination, namely percentage of sites allocated to the correct group, estimated by either the re-substitution method or the unbiased cross-validation method. However, this is not entirely appropriate for RIVPACS. In RIVPACS, test sites are not allocated to their most probable group (as done in most uses of discrimination), but rather are only allocated probabilistically. Thus the expected fauna for a test site is not just based on the observed fauna of the reference sites in the most probable group, but is a weighted average of the observed fauna of the reference sites in all of the TWINSPAN groups to which the discriminant functions indicate it has a probability of belonging. Outside of this project, we are currently trying to derive better general measures of the discriminatory effectiveness of different sets of variables in these situations. Although, this is beyond the scope of the current project, it may alter our conclusions about the relative effectiveness of different combinations of predictor variables and the use of GIS approaches to measuring some of them.

In summary, although altitude at site, distance from source, discharge category and slope at site are expected to be more precisely estimated using GIS than when obtained

manually from printed maps, the use of GIS values does not appear to give any improved ability to predict the biological group of a site, as assessed in tests using the RIVPACS reference sites, which cover the full range of environmental types of site available across Britain.

5. NEW RIVPACS PREDICTOR VARIABLES ACQUIRED FROM A GIS

5.1 Acquisition of New RIVPACS Predictor Variables from a GIS

5.1.1 Altitude of source, slope to source and stream power

The GIS methods for determining the source of a river site and the distance of the site from the source are outlined in Table 3.1. The altitude of the source is returned as the actual altitude point value at the source or the first altitude value downstream from the source. The latter is necessary in cases where sources do not have a underlying altitude point.

One principal use of deriving the altitude of the source is that it is then possible to estimate the average slope of the stream or river between the site and its source. This is referred to as 'slope to source' and defined as:

slope to source = ('altitude of source' – 'altitude of site') / 'distance to source'

where 'altitude of source', 'altitude of site' and 'distance to source' are all obtained from the GIS. The slope to source may provide a surrogate measure of the erosive power upstream of the site and hence provide a predictor of sediment type at the site.

A third new variable called 'unit' stream power (Ferguson, 1981) was defined as:

stream power = $g \cdot p \cdot Q \cdot S / W$

where g = gravitational acceleration = 9.81 m s^{-2}, p = density of water = 1000 kg m^{-3}, Q = discharge (m^3s^{-1}), S = stream slope at site (m km^{-1}), W = stream width (m). Stream power is a measure of the energy within a river system. The higher the stream power, the greater the potential to entrain large particles and to carry an increased sediment load. High stream power also increases the likelihood of an overall 'eroding' nature to the river environment (i.e. the site is a sediment source). Conversely low stream power increases the likelihood of a 'depositing' nature to the site environment (i.e. the site is a sediment 'sink'). Stream width (W) and slope (S) at the site are already RIVPACS variables. In RIVPACS discharge is recorded in logarithmic (doubling) categories, whereby discharge category 1 = < 0.31 m^3s^{-1}, 2 = 0.31-0.62 m^3s^{-1}, 3 = 0.62-1.25 m^3s^{-1}, 4 = 1.25-2.50 m^3s^{-1}, etc. Taking the mid-point of each category as the estimated discharge Q, estimates of stream power were derived for all the RIVPACS reference sites.

5.1.2 Upstream catchment area

Upstream catchment area was computed using the flow direction grid provided to CEH Dorset by CEH Wallingford. The flow-direction grid is based on 50 metre grid cells for the whole of mainland Britain. Each cell holds the direction of flow of water into it from the surrounding cells. From this, ArcView software procedures can compute the upstream catchment area (in hectares) from the number of cells which eventually flow into the cell containing the site. This process is completely independent of the blue-line river network and the final output is a polygon defining the upstream catchment area for

any site. Upstream catchment area increases with distance from source and there is a very close linear relationship on a log-log scale for most of the sites (Figure 5.1). However, using an automated procedure for determining the upstream catchment area based on the 50m resolution flow-direction grid appears to give grossly incorrect areas for some sites. Most noticeable are those 30-40 sites whose estimated upstream area is much less than is normal for sites that far from their source (Figure 5.1). In such cases, when the site's grid reference is snapped to the blue-line network, its position on the network places it in a 50m cell which is the flow accumulation cell for a tributary rather than the main river on which the site occurs. The estimated upstream catchment is then much smaller than its true area. Such mistakes could be corrected manually for the RIVPACS reference sites, but this would also need to be done manually for all other sites, such as the GQA sites. This is beyond the current scope of this project.

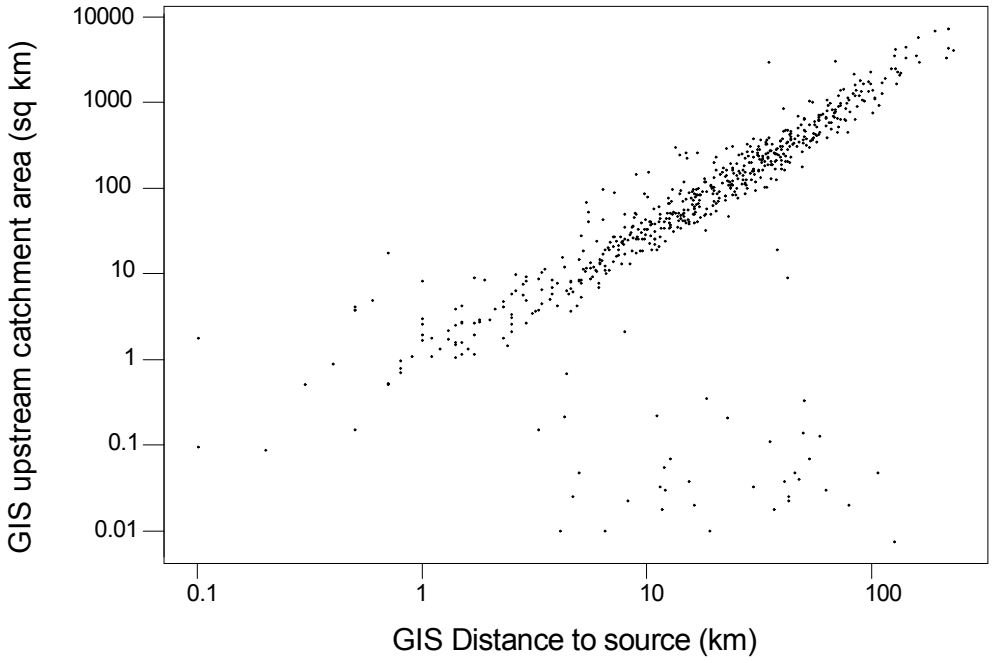


Figure 5.1: Relationship between the distance from a source of a suite and its upstream catchment area as estimated from the GIS using the GIS 50m resolution flow-direction grid

5.1.3 Site and upstream catchment geology

The surface (i.e. drift) and the underlying solid geology at and around a river site are likely to influence its physical character. Geology will also have some influence on the slope of the riverbed, the chemical composition of the water (e.g. pH, ionic composition, conductivity) and substrate composition at the site.

The British Geological Survey (part of NERC) have derived classifications of the solid geology and the drift geology of Britain. Both classifications are available in ArcView GIS format as GIS layers giving the predominant solid and drift geology class in each 1km square referenced to the National Grid. The solid geology consists of 115 BGS

classes and the drift geology of 13 classes (Tables 5.1 and 5.2). These very detailed classes provide too much detail and too many variables to be used directly within the RIVPACS discrimination. However, the BGS classes have been substantially reduced in number for use within the River Habitat Survey (Dawson, pers. comm.) and this simplified system of six River Habitat Survey (RHS) classes for solid geology and four RHS classes for drift geology (Tables 5.1-5.3) has been adopted here. The numeric codes and order for the classes correspond very roughly to increasing hardness of the geology.

Table 5.1: Solid Geology: BGS and RHS classes and descriptions

BGS Code	BGS description	RHS Code	RHS description
1	Undifferentiated gneiss	8	Hard Rocks
2	Metasediments	8	Hard Rocks
3	Marble	8	Hard Rocks
4	Anorthosite	8	Hard Rocks
5	Ultrabasic rock	8	Hard Rocks
6	Intermediate and basic rock	8	Hard Rocks
7	Gneissose granite, granite and pegmatite	8	Hard Rocks
8	Undifferentiated MOINE	8	Hard Rocks
9	Quartzite	8	Hard Rocks
10	Quartz-feldspar-granulite	8	Hard Rocks
11	Mica-schist, semi-pelitic schist and mixed schists	8	Hard Rocks
12	Granitic gneiss	8	Hard Rocks
13	Undifferentiated schist and gneiss of Shetland and Central Tyrone	8	Hard Rocks
14	Epidote-chlorite-schist, commanly hornblendic-Green Beds	8	Hard Rocks
15	Epidote-chlorite-schist, commanly hornblendic-Green Beds (Upper Dalradian)	8	Hard Rocks
16	Boulder bed and conglomerate	8	Hard Rocks
17	Quartzite grit, interstratified quartzose-mica-schist	8	Hard Rocks
18	Quartzose-mica-schist	8	Hard Rocks
19	Quartmica-schist, grit, slate and phyllite (Upper Dalradian)	8	Hard Rocks
20	Slate, phyllite and mica-schist	8	Hard Rocks
21	Slate, phyllite and mica-schist (Upper Dalradian)	8	Hard Rocks
22	Black shale with chert (Upper Dalradian)	8	Hard Rocks
23	Graphitic schist and slate	8	Hard Rocks
24	Limestone	8	Hard Rocks
25	Limestone (Upper Dalradian)	8	Hard Rocks
26	Serpentine	8	Hard Rocks
27	Epidiorite, hornblende-schist and allied types	8	Hard Rocks
28	Foliate granite, syenite and allied types	8	Hard Rocks
29	Hornblende schists	8	Hard Rocks
30	Gneiss, mica schists	8	Hard Rocks
31	Ultrabasic rock	8	Hard Rocks
32	Gabbro and allied types	8	Hard Rocks
33	Diorite and allied intermediate types	8	Hard Rocks
34	Granite, syenite, granophyre and allied types	8	Hard Rocks
35	Basalt dolerite, camptonite and allied types	8	Hard Rocks
36	Porphyrite, lamprophyre and allied types	8	Hard Rocks
37	Rhyolite, trachyte, felsite, elvans and allied types	8	Hard Rocks
38	Agglomerate in neck	8	Hard Rocks
39	Andesitic lava and tuff	8	Hard Rocks

BGS Code	BGS description	RHS Code	RHS description
40	Basalt, spilite and related tuff	8	Hard Rocks
41	Rhyolitic and trachytic lava and tuff undifferentiated	8	Hard Rocks
42	Basalt, spilite, hyaloclastic and related tuffs	8	Hard Rocks
43	Basaltic tuff	8	Hard Rocks
44	Andesitic lava and tuff, undifferentiated	8	Hard Rocks
45	Andesitic tuff	8	Hard Rocks
46	Rhyolitic lava	8	Hard Rocks
47	Ryolitic tuff, including ignimbrite	8	Hard Rocks
48	Tuff, undifferentiated, mainly andesitic	8	Hard Rocks
49	Basalt and spilite	8	Hard Rocks
50	Andesitic and basaltic lavas and tuffs, undifferentiated	8	Hard Rocks
51	Rhyolite, trachyte and allied types	8	Hard Rocks
52	Tuff (including ignimitbrite)	8	Hard Rocks
53	Basalt and spilite	8	Hard Rocks
54	Rhyolite, trachyte and allied types	8	Hard Rocks
55	Tuff, undifferentiated, mainly basaltic	8	Hard Rocks
56	Basalt	8	Hard Rocks
57	Basalt and spilite	8	Hard Rocks
58	Rhyolite, trachyte and allied types	8	Hard Rocks
59	Tuff, undifferentiated	8	Hard Rocks
60	Rocks of Anglesey, Lleyn Peninsular, Charnwood, Longmynd, etc	8	Hard Rocks
61	Sandstone and grit	5	Sandstone
62	Pipe-Rock and Basal Quartzite	8	Hard Rocks
63	Serpulite Grit and Fucoid Beds	8	Hard Rocks
64	Lower CAMBRIAN	8	Hard Rocks
65	Middle CAMBRIAN	8	Hard Rocks
66	Upper CAMBRIAN, including Tremadoc	8	Hard Rocks
67	Durness Limestone (partly Cambrian)	7	Limestone
68	Llanvirn and Arenig	8	Hard Rocks
69	Llandeilo	8	Hard Rocks
70	Caradoc	8	Hard Rocks
71	Ashgill	8	Hard Rocks
72	Llandovery	8	Hard Rocks
73	Wenlock	7	Limestone
74	Ludlow	7	Limestone
75	Lower Old Red Sandstone, including Downtonian	5	Sandstone
76	Lower Devonian (England and Wales only)	5	Sandstone
77	Middle Old Red Sandstone (Scotland) Middle Devonian (England)	5	Sandstone
78	Upper Old Red Sandstone (Scotland) Upper Old Red Sandstone and Upper Devonian (England)	5	Sandstone
79	Basal Conglomerate (including possible Devonian)	5	Sandstone
80	Tournaisian and Visean (Carboniferous Limestone Series)	6	Chalk
81	Namurian (Millstone Grit Series)	5	Sandstone
82	Lower Westphalian (mainly Productive Coal Measures)	7	Limestone
83	Upper Westphalian (including Pennant Measures)	7	Limestone
84	Westphalian and ?Stephanian, undivided, of Barren Red lithology	7	Limestone
85	Permian basal breccias, Sandstones and mudstones	5	Sandstone
86	Magnesian Limestone (Permian)	7	Limestone
87	Permian mudstones (including Middle and Upper Marls, Eden ans St Bees shales)	4	Shale
88	Budleigh Salterton Pebble Beds	5	Sandstone
89	Permian and Triassic Sandstones, undifferentiated, including Bunter and Keuper	5	Sandstone

BGS Code	BGS description	RHS Code	RHS description
90	Triassic mudstones including Keuper Marl, Dolomitic Conglomerate and Rhaetic)	4	Shale
91	Lower Lias	3	Clay
92	Middle Lias	3	Clay
93	Upper Lias	3	Clay
94	Inferior Oolite	6	Chalk
95	Great Oolite	6	Chalk
96	Cornbrash	6	Chalk
97	Oxford Clay and Kellaways Beds	3	Clay
98	Corallia	3	Clay
99	Kimmeridge Clay and Ampthill Clay	3	Clay
100	Portland Beds	6	Chalk
101	Purbeck Beds	6	Chalk
102	Hastings Beds	3	Clay
103	Weald Clay	5	Sandstone
104	Lower Greensand	5	Sandstone
105	Upper Greensand and Gault (England) 8Greensand (Scotand)	5	Sandstone
106	Upper Chalk (Scotland) Chalk including Red Chalk (England)	6	Chalk
107	Inter-lava beds (Scotland) Oldhaven, Blackheath, Woolwich, & Reading & Thane	3	Clay
108	Inter-lava beds (Scotland) London Clay (England)	3	Clay
109	Inter-lava beds (Scotland) Barton, Bracklesham and Bagshot Beds (England)	5	Sandstone
110	Lough Neagh Clays (Scotland) Bovey Formation, St Angus Sands, etc (England)	3	Clay
111	Hampstead Beds and Bembridge Marls	NIL	---
112	Gravel (Scotland) Lenham Beds (England)	3	Clay
113	Gravel (Scotland) Coralline Crag (England)	3	Clay
114	Gravel (Scotland) St Erth Beds Cornwall))	3	Clay
115	Norwich Crag, Red Crag and Chillesford Clay	3	Clay

Table 5.2: Drift geology: BGS and RHS classes

BGS code	BGS description	RHS Code	RHS description
0	No Drift Geology	0	No drift geology
1	Landslip	5	Sandstone
2	Blown Sand	5	Sandstone
3	Peat	1	Peat
4	Lacustrine Clays, Silts and Sands	3	Clay
5	Alluvium	2	Alluvium
6	River Terrace deposits (mainly sand and gravel)	5	Sandstone
7	Raised Beach and Marine deposits	5	Sandstone
8	Glacial Sand and Gravel	5	Sandstone
9	Boulder Clay and Morainic drift	3	Clay
10	Sand and Gravel of uncertain age or origin	5	Sandstone
11	Clay with flints	3	Clay
12	Brickearth, mainly loess	3	Clay
13	Crag	3	Clay

Any impacts of upstream catchment geology on the macroinvertebrate fauna at a site are likely to decrease with distance from the site. The geology in the immediate vicinity of the site is likely to have the most impact on stream bed characteristics, but water

chemistry may be more influenced by the whole upstream geology. We have derived two sets of geological variables. The first set estimates the percentage of the total upstream catchment area occupied by each RHS class of solid geology and by each RHS class of drift geology. The second set are presence-absence variables denoting the solid and drift RHS class for the 1 km square in which the site occurs. Although almost all areas and 1 km squares can be assigned a solid geology class, over 44% of the RIVPACS reference sites are in 1 km squares where either there is no drift geology or no class has been recorded (Table 5.3).

Table 5.3: RHS classes of solid and drift geology; the percentage of reference sites in 1 km squares of each class are given

Solid geology			Drift geology		
code	description	% of sites	code	description	% of sites
0	No RHS class	1.0	0	No drift geology	43.8
3	Clay	15.1	1	Peat	2.3
4	Shale	4.6	2	Alluvium	22.8
5	Sandstone	32.2	3	Clay	18.2
6	Chalk	21.0	5	Sandstone	12.9
7	Limestone	4.9			
8	Hard rocks	21.2			

5.1.4 Stream order

Stream order was computed for the 1:50,000 scale river network as defined by Strahler (1957). An algorithm described by Lanfear (1990) for automatically computing Strahler stream order from vector networks was adapted to run in the GIS software ArcView. Hydrometric areas were processed one at a time and the stream order was attached to each arc as an attribute. The algorithm was capable of handling braided streams. However, stream orders computed for arcs in flat lowland areas with grid-like drainage sections were meaningless. As these sections had already been labelled as "not traceable" within the CEH river network GIS, this was not a concern.

It should be noted that the stream order of the arc representing the mouth of each river is highly dependent on the presence or absence of first order streams. Apparently, Ordnance Survey can and on occasions do omit watercourses from maps to reduce clutter and enhance clarity. As the 1:50,000 network was generated from OS data it can be assumed that some watercourses do not attain their true stream order value on the digital network. However, there is no easy solution to this dilemma and the 1:50,000 Landranger series tends to be the map of choice for nearly all research in ecology, including any determination of stream order.

As would be expected, Strahler stream order is highly correlated with distance from source with a rank correlation of 0.84 (Table 5.4, Figure 5.2). This may limit its ability to improve our prediction of the TWINSPAN site group of the reference sites.

Table 5.4: Distance from source of the 614 RIVPACS reference sites classified by their Strahler stream order

Stream order	n	Median	25%	75%	Min	Max
1	41	1.4	0.7	3.0	0.1	41.0
2	73	4.0	1.5	6.0	0.3	19.0
3	117	9.0	5.0	14.5	1.0	45.0
4	156	17.0	10.0	25.9	1.9	63.6
5	130	32.0	20.7	44.0	7.0	133.0
6	79	55.0	38.9	80.0	18.0	202.8
7	18	97.8	77.5	138.5	51.0	202.8

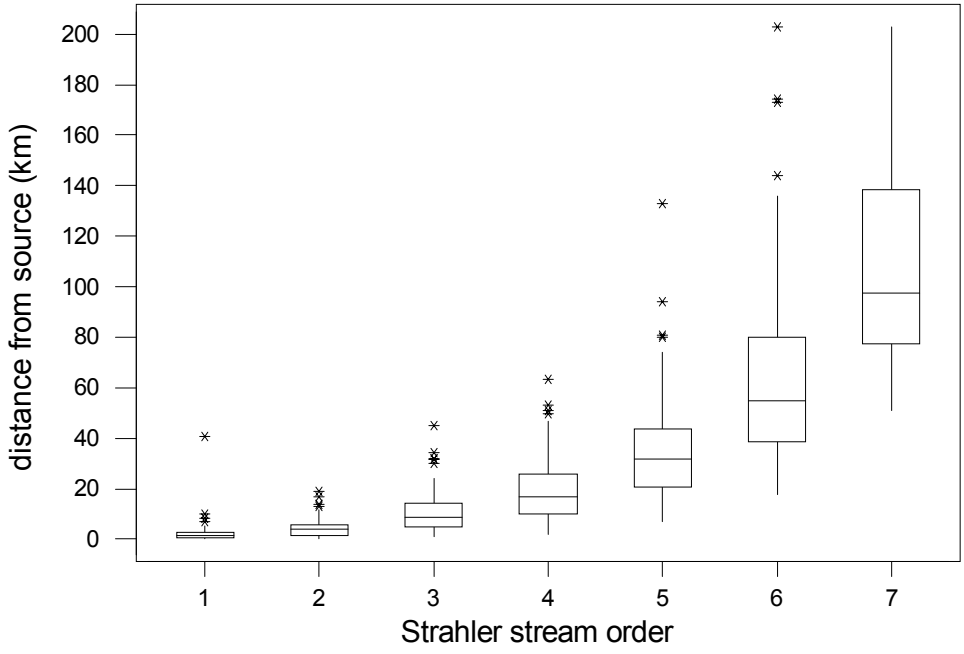


Figure 5.2: Boxplot summarising the variation in distance from source of the RIVPACS reference sites in relation to their Strahler stream order. Boxplot definition: box shows inter-quartile range (25-75%) with median (50%) shown as horizontal line; outer range (min-max) shows as vertical line except for individual outlier observations marked as *

5.2 Effectiveness of Using Existing and New Variables Derived from a GIS in RIVPACS Predictions

5.2.1 Geological variation between reference site groups

The percentage of the upstream catchment area which is in each RHS solid geology class varies in a systematic pattern between the TWINSPAN groups of RIVPACS reference sites (Figures 5.3 and 5.4, Table 5.5). In particular, site groups 1, 2, 10-12, 14-15 and 18 are dominated by underlying hard rocks upstream. Site groups 8, 25, 27 and 30-33 are predominantly underlain by chalk (Table 5.5). The highest proportions of

upstream catchment areas with clay solid geology occur mostly in site groups 5, 8, 9 and all the lowland stream and river site groups 25-35, with the exception of site group 28 whose sites tend to be downstream of shale rather than clay catchments (Table 5.5). The differences between the TWINSPAN site groups in terms of the percentage of the upstream catchment area covered by each RHS drift geology class are summarised in Figure 5.5 and Table 5.5. Just under half (44%) of all sites either have no drift geology or have none recorded. Peat is the dominant RHS drift geology class in the upstream catchments of sites in groups 1, 2, 12, 13 and 34, whilst 'sandstone' drift class is most common for site groups 9 and 19 (Table 5.5). There are upstream catchment areas whose RHS drift geology class is clay for some reference sites in every TWINSPAN group except 5 and 9, with the average percentage cover ranging from 7 to 47% (Table 5.5). Alluvium drift generally covers between 1 and 10% of the upstream area of sites in most groups (Figure 5.5).

The influence of solid and drift geology type in the vicinity of the site (i.e. the 1 km square in which it resides) was also assessed. Table 5.6 shows the percentage of sites in each TWINSPAN group whose underlying geology falls in each RHS solid and drift class. As would be expected, a considerable proportion of sites are underlain by the alluvium drift class as they are in current or historic floodplains. Site group 34 is unusual in that 62% of its 13 reference sites have peat drift geology (Table 5.6).

5.2.2 Individual effectiveness of new GIS variables

To assess whether any of the GIS variables improved the prediction of the expected macroinvertebrate fauna for the reference sites, the GIS variables were first added one at a time (Table 5.7). As already reported in section 4 (Table 4.1), none the GIS versions of the original RIVPACS variables, altitude at site, distance from source, slope at site or discharge category gave any improvement over using the current manually estimated versions of the variables.

The two new variables, altitude at source and the average slope between the site and its source, had as high a discriminatory ability as most of the original RIVPACS variables when used individually, but neither variable improved predictive ability when added to the original RIVPACS variables, as judged by the cross-validation results. However, adding the derived variable, referred to as 'stream power', gave minor improvements in percentage prediction to correct TWINSPAN group (Table 5.7).

Upstream catchment area (used in logarithmic form) was not as good a predictor of TWINSPAN group as distance from source and did not improve current predictive ability in RIVPACS III+. Similarly, Strahler stream order failed to increase the percentage of sites predicted to correct group, based on the cross-validation procedure (Table 5.7).

Although there were systematic differences between the TWINSPAN groups of reference sites in their upstream geology, adding the variables representing the proportion of the upstream catchment with each of the RHS major classes of solid and drift geology did not seem to give any improvement in prediction of the RIVPACS biological site groups (Table 5.7). Nor did the solid or drift class of the 1km square in which the site lay help improve predictive ability.

It must be remembered that for a small proportion of sites, the incorrect upstream catchment has probably been determined by the automatic GIS procedures described in section 5.1.2. Therefore the estimated upstream catchment geological character for such sites is also likely to be incorrect. Predictive ability might improve if all such errors were eliminated.

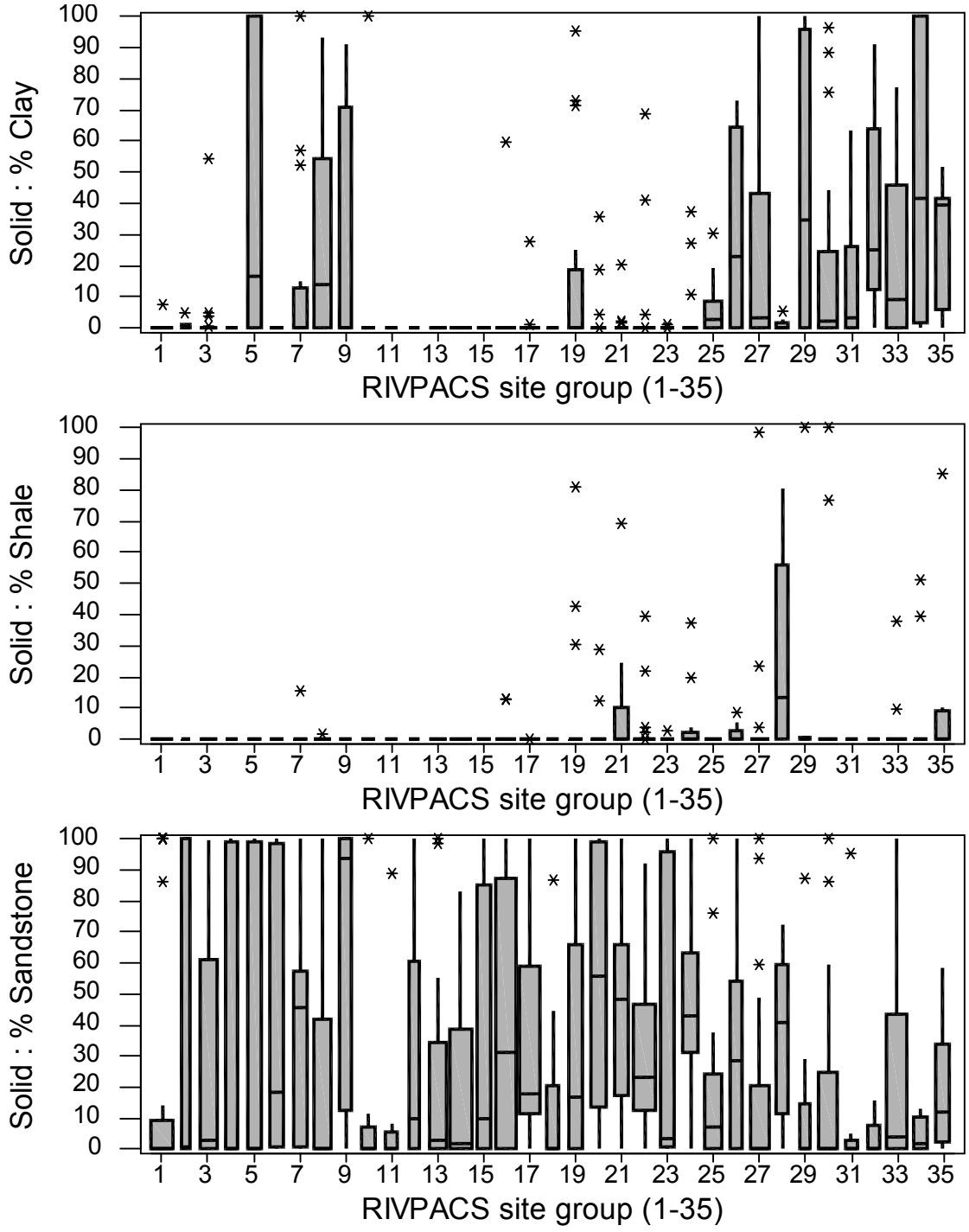


Figure 5.3: Boxplots of the percentage of the upstream catchment area for which the underlying solid geology is RHS class "Clay", "Shale" or "Sandstone", given separately for the reference sites in each of the 35 TWINSPAN groups. Boxplot definition as in Figure 5.2

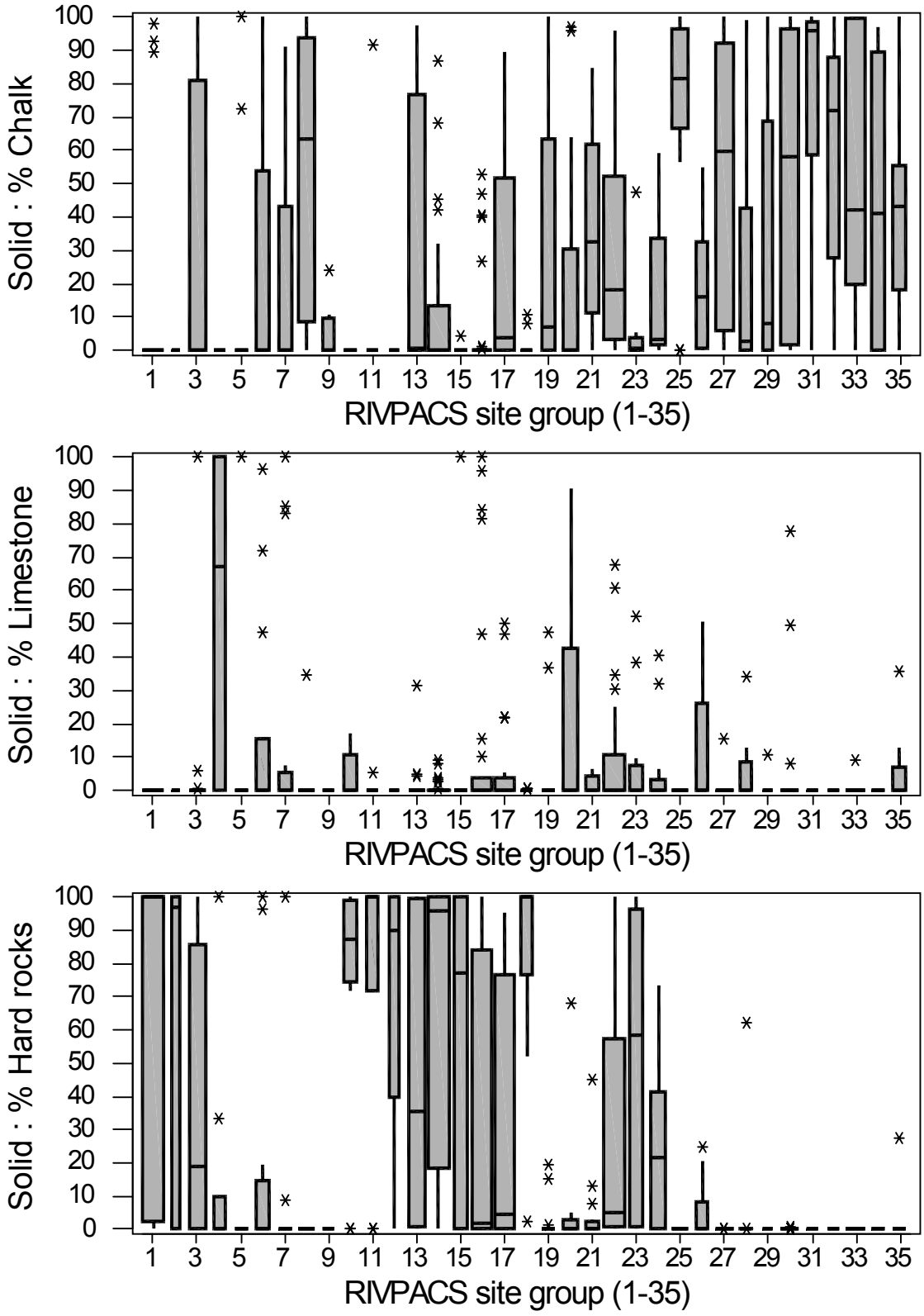


Figure 5.4: Boxplots of the percentage of the upstream catchment area for which the underlying solid geology is RHS class "Chalk", "Limestone" or "Hard rocks", given separately for the reference sites in each of the 35 TWINSPAN groups. Boxplot definition as in Figure 5.2

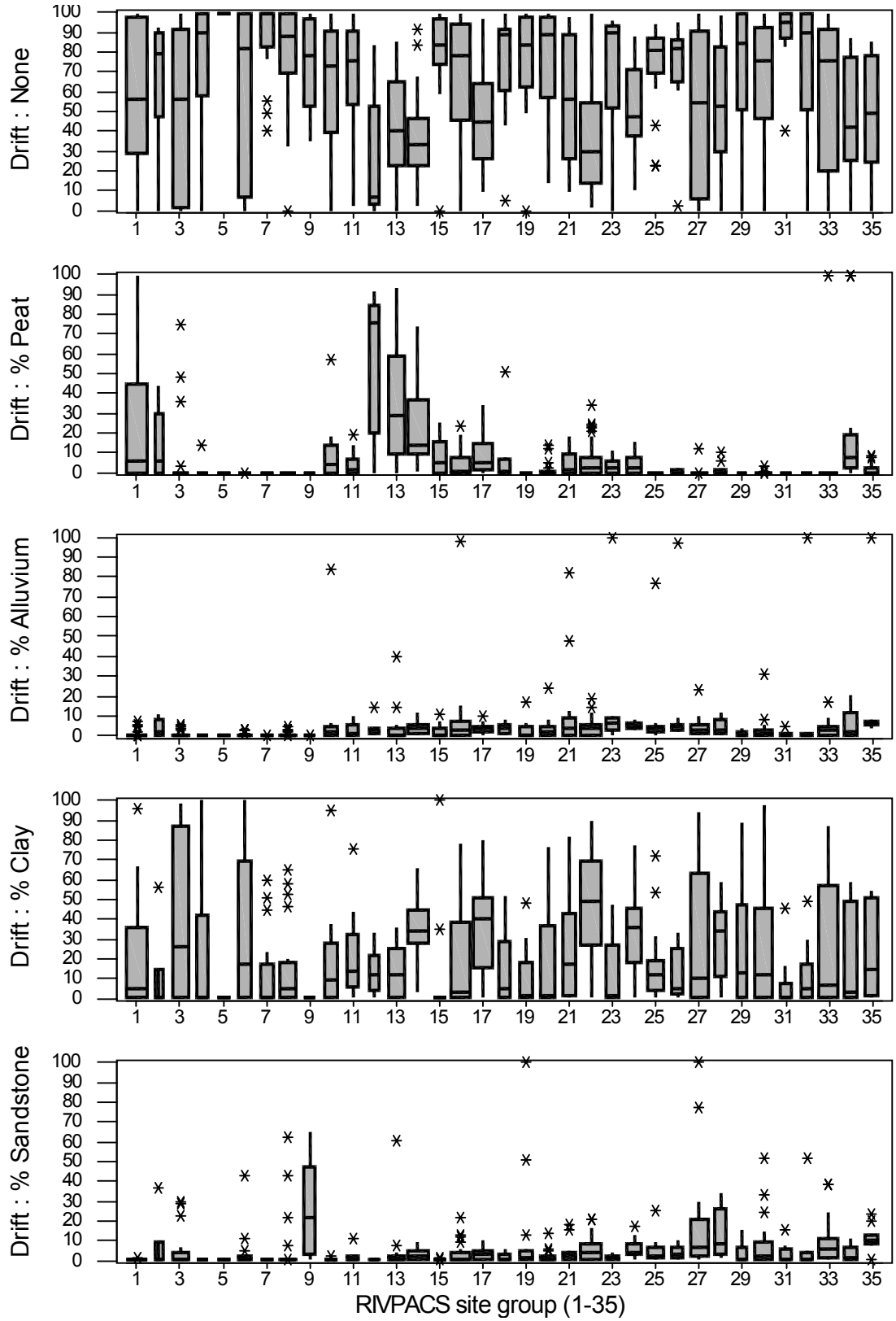


Figure 5.5: Boxplots of the percentage of the upstream catchment area for which the underlying drift geology is RHS class "None", "Peat", "Alluvium", "Clay" or "Sandstone", given separately for the reference sites in each of the 35 TWINSPAN groups. Boxplot definition as in Figure 5.2

Table 5.5: Average percentage of the upstream catchment area in each RHS solid and drift geology class, separately for the RIVPACS reference sites in each TWINSPAN group. Zero values are omitted for clarity

Site	RHS Solid geology class						RHS Drift geology class			
group	Clay	Shale	Sandstone	Chalk	Sandstone	Hard rocks	Peat	Alluvium	Clay	Sandstone
1			21	8		71	25	1	18	
2	1		34			66	14	4	9	6
3	3		27	29	5	35	8	1	39	5
4			34		52	13	1		22	
5	44		33	14	8					
6			42	26	16	17			35	4
7	14	1	39	21	17	7			11	
8	28		18	52	2			1	14	6
9	24		68	4						25
10	8		10		4	76	11	8	17	
11			10	9	1	80	5	3	22	2
12			29			71	60	4	13	
13			21	30	2	46	35	4	13	4
14			18	11	1	70	23	4	34	2
15			32		8	59	8	2	11	
16	2	1	41	7	14	35	4	7	17	2
17	1		34	25	6	33	9	3	37	3
18			13	1		85	6	3	13	1
19	17	10	34	29	5	2		3	9	11
20	3	2	54	17	19	4	2	3	15	2
21	2	8	48	37	2	4	5	12	23	4
22	3	2	29	29	8	29	6	4	47	5
23			39	4	7	48	4	12	12	1
24	4	4	48	17	5	21	4	5	33	6
25	5		18	75				7	15	4
26	31	2	34	18	11	5	1	12	10	3
27	25	5	17	53	1		1	4	30	15
28	1	28	36	24	6	6	2	4	30	12
29	45	11	13	29	1			1	26	3
30	18	7	16	50	6			3	24	7
31	13		10	77				1	7	3
32	34		3	63				10	11	6
33	23	2	24	51			3	3	24	8
34	44	7	5	44			22	6	21	3
35	26	9	19	39	5	2	2	13	23	11

Table 5.6: Percentage of the RIVPACS reference sites in each TWINSPAN group which lie in 1 km squares dominated by each RHS solid and drift geology class. Zero values are omitted

Site	RHS Solid geology class						RHS Drift geology class			
group	Clay	Shale	Sandstone	Chalk	Sandstone	Hard rocks	Peat	Alluvium	Clay	Sandstone
1			18	9		74	6	9	32	
2			33			67	17	17	17	
3			30	35	10	25		10	30	20
4			27		36	27			9	
5	50		33	8	8					
6			43	29	14	14			36	14
7	31	6	38		19	6			6	
8	45		23	32				5	23	
9	20		70							10
10	8					85		31	15	8
11			30	10		60		30	10	20
12			25			75		63	25	13
13			20	25		55	5	30	35	10
14			22	16		59		38	25	16
15			58		8	33	8	17	8	
16		3	45	19	13	19		13	19	19
17	4		43	29	7	18		32	43	18
18			15	8		77	8	8	8	15
19	19	13	38	25				19	6	6
20		10	40	25	25			5	10	10
21	13	19	56	13				50	19	
22	3	8	36	28	8	18		33	31	21
23			53		13	33		40	20	
24	6	12	53	24	6			41	18	24
25	24		29	48				24	10	29
26	17	8	58	17				25	17	8
27	32	8	16	44				36	8	20
28	20	40	20	20				30	20	20
29	56	11	22	11					11	22
30	33	8	25	29				17	17	17
31	30		20	50				10		10
32	50		20	30				20	10	
33	29	6	32	32				35	6	23
34	69	8	15	8			62	31		
35	36	7	36	21				50	14	36

Table 5.7: Ability of each new GIS variable to predict the TWINSPAN biological group of the 614 RIVPACS reference sites, (a) when used on its own, and (b) to improve predictions over just using the current RIVPACS environmental variables. * denotes using new GIS rather than original version of variable

New variables used	(a) used on their own		(b) used in addition to current RIVPACS variables (option 1)	
	cumulative %classified to correct group by:		cumulative %classified to correct group by:	
	re-substitution	cross-validation	re-substitution	cross-validation
None			51.3	41.2
Log distance from source*	13.2	12.7	52.8	40.4
Log slope (at site)*	14.3	13.7	52.1	40.2
Discharge category*	12.2	12.2	51.3	39.7
Log altitude (at site)*	10.0	9.8	51.1	38.4
Log altitude at source	14.0	13.5	52.1	41.0
Log slope to source	14.5	13.8	53.1	41.4
Log stream power	11.6	11.4	53.3	41.9
Log catchment area	9.8	9.6	51.8	41.1
Stream order	12.9	12.9	52.9	40.7
% upstream catchment area in each of:				
6 RHS solid geology classes	13.0	10.9	51.5	39.6
4 RHS drift geology classes	11.6	11.1	52.3	40.6
Geology class of 1km square of the site:				
Solid geology class	10.9	6.2	51.3	37.6
Drift geology class	8.8	3.3	53.3	38.6
All of the above GIS variables			56.8	37.6

In summary, automated GIS procedures have been developed to derive new environmental predictor variables for any river site. These are: altitude at source, average slope from site to source, a measure of stream power, upstream catchment area, the proportion of the upstream catchment covered by each major RHS solid and drift geology class and the solid and drift geology class of the 1km square containing the site. When assessed on the RIVPACS reference sites, none of the GIS variables currently provides any significant improvement in ability to predict site group and hence expected fauna when used with the existing RIVPACS predictor variables. There was only some minor improvement when adding stream power (a function of discharge, stream width and slope at site).

However, the automated GIS procedures identify an incorrect upstream catchment for a small proportion of sites (ca. 5-10%); developing GIS procedures to "manually" correct such site positions is beyond the scope of this project, but merits further investigation.

6. CONCLUSIONS AND RECOMMENDATIONS

6.1 Conclusions

Previous sections of this report have demonstrated that it is possible to acquire some of the existing and also some new environmental variables for prediction from a Geographic Information System (GIS). This has entailed the development of novel and sophisticated procedures for the efficient automated or semi-automated extraction and collation of the required data. The work necessary to develop this system has been much greater and more time-consuming than envisaged when this package was drafted at the outset of the project.

Acquisition of some variables for prediction such as distance from source and altitude of site using a GIS is generally more reliable than acquisition of these same features from a map. In addition, the GIS delivers results for both existing and new variables more rapidly than was possible using map-based procedures. Although a batch mode of operation is possible using GIS, at present, it is advisable to deal with each site interactively and confirm its true location on the blue line network to avoid the possibility of the site grid reference 'snapping' to an incorrect river/tributary.

However, a number of problems were uncovered during this research, and not all of them have been fully resolved. For example, in some limited low-lying areas of England, it has not been possible to develop automated procedures for locating the source of a stream due to the grid-like drainage patterns encountered in such areas. Also, when generating the slope at a site using the GIS, a different procedure was used to that employed in the original map-based method for RIVPACS. There were also practical problems in using the discharge category layer in the GIS because the layer supplied was at a different scale and lower resolution than the blue line data.

This research has highlighted the need for the geographic position of biological sampling sites within the GIS to be recorded to a higher level of accuracy than the 100 m resolution used at present. Ideally the geographic position of sites should be identified as the required point on the blue-line network (i.e. to the nearest metre), so as to avoid the biological sampling site being associated with the wrong stretch of river or tributary.

Hence, a question remains about when it will be appropriate to take advantage of these GIS developments. Any changes to the present RIVPACS system should only be made when there is certainty that the site values for environmental variables generated by GIS are as accurate or more accurate than those derived by other means. It is also important to recognise that it would be inappropriate to make a series of piecemeal changes because each change has numerous ramifications for RIVPACS itself and for future predictions at test sites.

Although the ease of acquisition of current and new predictor variables for RIVPACS is an incentive for using GIS-derived data at some point in the future, a further substantial boost would be clear evidence that GIS-derived variables lead to an increase in the predictive capability of the system. At present, the evidence for this is lacking, or at least equivocal. It now seems unlikely that any substantial increase in predictive

capability is possible, given the range of variables which have been examined in previous phases of RIVPACS development (Wright *et al.,* 1984) and during the current exercise. This does not, however, exclude the possibility that a future version of RIVPACS, with additional sites and an entirely new classification system, could show increased predictive capability based on current and new GIS-derived variables.
When the time comes to consider implementing the use of GIS-derived data in predictions, based on an improved/acceptable value for the percentage of sites predicted to the correct group on an internal test of the RIVPACS reference sites, it will also be necessary to do some further tests before recommending the adoption of a new set of variables as the standard. For example, it will be important to undertake predictions of BMWP indices, families and species to ensure that the distributions of O/E ratios for the reference sites is as close (or closer) to unity than those achieved for the Option 1 variables.

Assuming that the evidence is convincing that the use of GIS-derived predictor variables is beneficial, then the following steps can be envisaged:

1. Use predictor variables (including GIS-derived variables where appropriate) to derive new MDA equations for the RIVPACS reference sites for Great Britain. This could be undertaken on the existing 614 sites or a modified set of reference sites, depending on the timing of the work (see Section 5.7.2 in the associated Stage 3 E1-007/TR report for the range of future potential developments).

2. Use the new MDA equations to derive a new version of RIVPACS. Produce an updated manual and issue the new software.

3. Ensure that all those using the new prediction system, and especially the Environment Agency, have ready access to GIS-derived site variables for use at all their test sites.

4. Investigate the relationship between, for example, the O/E (= EQI) values for the 6000+ GQA sites, derived from the present system and the new GIS-based system.

As indicated in the acknowledgements section of this report, funding from the Environment Agency for blue line GIS-development has been supported by the River Habitat Survey (RHS) project in addition to this package. In future, work undertaken under RHS or other projects is likely to see further developments to the system. However, within this project it is necessary to report on the progress made within the cost and time constraints of the RIVPACS package. At the present time, it would be premature to offer outline costs for using GIS-derived site variables together with the consequences for RIVPACS itself and RIVPACS predictions as outlined above.

6.2 Recommendations

As stated in the Stage 3 report E1-007/TR (Section 5.7.2), there is a wide range of potential developments envisaged for RIVPACS as a result of the current contract, which includes 10 separate packages. Therefore, it is crucial that the Environment

Agency takes a strategic view at an early stage, in consultation with CEH Dorset, to ensure that all developments are undertaken in a logical manner.

Recommendation 1. ***We believe that the most urgent need is for a Windows version of RIVPACS before further changes are made to the existing software.***

In the previous section we indicated that very substantial progress has been made in the acquisition of site variables from a GIS. Already, these are being exploited within the River Habitat Survey. However, we also expressed caution over the premature use of GIS-derived site variables in RIVPACS predictions because of the many consequences which flow from this decision.

Recommendation 2. ***Our judgement is that it would be best to defer use of GIS-derived site variables within RIVPACS until further development work has taken place and there is clear evidence that the GIS system can deliver reliable outputs throughout Great Britain for a number of specified variables.***

Clearly, the rate at which progress can be made will depend on the financial support available to progress the work. It is for the Environment Agency to decide whether it wants to fund additional development work under RIVPACS or wait for progress via other routes.

Recommendation 3. ***The Environment Agency to consider whether it wants to fund additional development work on the extraction of predictor variables from a GIS and subsequent assessment of their value within RIVPACS.***

REFERENCES

Clarke R T, Furse M T and Wright J F, 1994. *Testing and further development of RIVPACS. Phase II: Aspects of robustness*. IFE Interim report (R&D 243/7/Y). National Rivers Authority, Bristol.

Clarke R T, Furse M T, Wright J F, Moss D, 1996. Derivation of a biological quality index for river sites: comparison of the observed with the expected fauna. Journal of Applied Statistics, 23: 311-332.

Ferguson R I, 1981. *Channel form and channel changes*. In: British Rivers (Ed. J. Lewin). George Allen & Unwin, London, pp 90-125.

Furse M T, 2000. The application RIVPACS procedures in headwater streams – an extensive and important national resource. In: J.F. Wright, D.W. Sutcliffe and M.T. Furse (Editors). Assessing the Biological Quality of Freshwaters: RIVPACS and Other Techniques, Freshwater Biological Association, Ambleside, pp 79-92.

Lanfear K J, 1990. *A fast algorithm for automatically computing Strahler stream order*. Water Resources Bulletin, **26**, 6: 977 – 981.

Moss D, Furse M T, Wright J F, Armitage P D, 1987 The prediction of the macro-invertebrate fauna of unpolluted running-water sites in Great Britain using environmental data. Freshwater Biology, **17**: 41-52.

Murray-Bligh J A D, Furse M T, Jones F H, Gunn R J M, Dines R A, Wright J F, 1997. Procedure for Collecting and Analysing Macroinvertebrate Samples for RIVPACS. Joint publication by the Institute of Freshwater Ecology and the Environment Agency, 162 pp.

SAS, 1989. SAS/STAT User's Guide, Version 6, 4th edition, Vol.2., SAS Institute, Cary, 1686 pp.

Strahler A N, 1957. Quantitative analysis of watershed geomorphology. Transactions of the American Geophysical Union 38: 913 – 920.

Wright J F, Moss D, Armitage P D and Furse M T, 1984. A preliminary classification of running-water sites in Great Britain based on macroinvertebrate species and the prediction of community type using environmental data. *Freshwater Biology*, 14, pp. 221-256.

Wright J F, 2000. An introduction to RIVPACS. In: J F Wright, D W Sutcliffe and M T Furse (Editors) Assessing the Biological Quality of Freshwaters: RIVPACS and Other Techniques, Freshwater Biological Association, Ambleside, pp 1-24.

Wright J F, Winder J M, Clarke R T and Davy-Bowker J, 2002. Testing and Further Development of RIVPACS, Stage 3. R&D Technical Report E1-007/TR. Environment Agency, Bristol.