EA-Anglian Box 14

(EA)

[redacted]

ENVIRONMENT AGENCY

Anglian Regional Operational Investigation 579

Quantification of the relationship between effluent quality and biological quality

Final Report

WRc Ref: CO 4261/1
December 1996

RESTRICTED

ENVIRONMENT AGENCY

NATIONAL LIBRARY &
INFORMATION SERVICE

ANGLIAN REGION

Kingfisher House, Goldhay Way,
Orton Goldhay,
Peterborough PE2 5ZR

ENVIRONMENT AGENCY
124320

QUANTIFICATION OF THE RELATIONSHIP BETWEEN EFFLUENT QUALITY AND BIOLOGICAL QUALITY: DRAFT FINAL REPORT.

A Gunby, I Milne and M Wheeler

Research Contractor:
WRc plc
Henley Rd Medmenham
Marlow
SL7 2HD

Environment Agency, Anglian Region

Environment Agency, Anglian Region
Kingfisher House
Golday Way
Orton Golday
Peterborough
PE2 5ZR

Tel: 01733 371811
Fax: 01733 231840

© Environment Agency 1997

All rights reserved. No part of this document may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise without the prior permission of the Environment Agency.

The views expressed in this document are not necessarily those of the Environment Agency. Its officers, servants or agents accept no liability whatsoever for any loss or damage arising from the interpretation or use of the information, or reliance upon views contained herein.

Dissemination status
Internal: Limited Release
External: Restricted

Statement of use
This report is for use by water environment practitioners with involvement in water quality assessment.

Research contractor
This Research Contractor for this work was:

WRc plc
Henley Rd Medmenham
Marlow
SL7 2HD

WRc Report Nº CO4261/1/09106

Environment Agency Project Leader
The Environment Agency's Project Leader for Regional Operational Investigation 579 is:

Dr Sarah Chadd - Environment Agency, Anglian Region, Peterborough

Further Information:
Further information relating to this document may be obtained from the Anglian Region R&D Management Support Officer in our Peterborough Office.

CONTENTS Page

LIST OF TABLES ii

LIST OF FIGURES ii

EXECUTIVE SUMMARY 1

KEY WORDS 1

1. INTRODUCTION 3

2. THE DATASET 5

2.1 Data received 5
2.2 Manipulation of data and construction of database 5
2.3 Original dataset 7
2.4 Augmented dataset 8

3. ASSESSMENT OF IMPORTANT FACTORS 9

3.1 Introduction 9
3.2 Biological determinands 10
3.3 Effluent chemistry determinands 10
3.4 Examination of possible relationships 11

4. IDENTIFICATION OF RELATIONSHIPS BETWEEN EFFLUENT QUALITY AND BIOLOGICAL QUALITY 21

4.1 Introduction 21
4.2 Principal coordinates analysis 22
4.3 Canonical correlation 27

5. DISCUSSION AND CONCLUSIONS 31

6. RECOMMENDATIONS 33

REFERENCES 35

Page

LIST OF TABLES

Table 2.1	Number of STWs in original database with associated biological monitoring sites	7
Table 2.2	Number of STWs in original database with data at associated upstream and downstream sites	7
Table 2.3	Number of STWs in original database with all associated data	8
Table 4.1	Percentage variation explained by the principal coordinates of the biological changes variables	23
Table 4.2	Percentages of the total variation explained by the principal coordinates of the effluent variables.	23
Table 4.3	Percentages of the total variation explained by the principal coordinates of the upstream GQA variables.	24
Table 4.4	Percentages of the total variation explained by the principal coordinates of the upstream biology variables.	24
Table 4.5	Percentages of the total variation explained by the principal coordinates of the upstream RIVPACS predictions variables.	25
Table 4.6	Analysis of variance table for the regression	25
Table 4.7	Estimates of regression coefficients (STW effect levels are not shown)	26
Table 4.8	Accumulated analysis of variance: contributions to the sums of squares from adding each explanatory variable to the model in the order listed (+ sign), and then dropping them in the order listed (- sign).	26
Table 4.9	Canonical correlations	28
Table 4.10	Coefficients of variables in relationships	28

LIST OF FIGURES

Figure 3.1	Relationship between BMWP EQI ratio and theoretical BOD increase	12
Figure 3.2	Relationship between ASPT EQI ratio and theoretical BOD increase	12
Figure 3.3	Relationship between BMWP EQI and theoretical ammonia increase	13
Figure 3.4	Relationship between ASPT EQI and theoretical ammonia increase	13
Figure 3.5	Relationship between BMWP EQI ratio and change in GQA BOD	14

Page

Figure 3.6 Relationship between ASPT EQI ratio and change in GQA BOD 15

Figure 3.7 Relationship between BMWP EQI ration and change in GQA ammonia 15

Figure 3.8 Relationship between ASPT EQI ratio and change in GQA ammonia 16

Figure 3.9 Relationship between BMWP EQI ratio and change in GQA dissolved oxygen 16

Figure 3.10 Relationship between ASPT EQI ratio and change in GQA dissolved oxygen 17

Figure 3.11 Relationship between GQA change and theoretical increase for BOD 18

Figure 3.12 Relationship between GQA change and theoretical increase for ammonia 18

Figure 3.13 Theoretical increase in BOD plotted against time 19

Figure 3.14 GQA BOD change plotted against time 20

Figure 3.15 Ratio of ASPT EQI plotted against time 20

EXECUTIVE SUMMARY

This report is the final output from the Environment Agency, Anglian Region Operational Investigation 597: Quantification of the relationship between effluent quality and biological quality.

The Agency has a national system for biological assessment of water quality but no formal method of relating discharge quality to biologically assessed water quality. This project arose from an identified need for such a methodology, which would enhance the Agency's ability to target investment and demonstrate improvements.

The objectives of the project were, firstly to use existing data to identify relationships between the chemical quality of sewage treatment works (STW) effluent and the biological quality of receiving waters, and, secondly, to use these relationships to develop a predictive protocol for assessing the likely effects on biological quality of effluent improvements.

Data on a wide range of appropriate variables was provided and a series of data manipulations and multivariate statistical analyses were carried out to try and identify relationships of interest.

Unfortunately it was not possible to identify any useful relationships, primarily because the majority of the STWs for which all the necessary variables were available were not actually having a large impact on the receiving waters (at least according to the data used) and also because of the high degree of variability in the data, particularly the biological data.

The failure to find useful relationships, meant that it would not be possible to develop the envisaged protocol and the project was therefore terminated. This report documents the data analysis exercise undertaken.

From the outset of this project it was recognised that an experimental approach, with targeted sampling, might be required but that existing data should first be assessed to avoid unnecessary effort. The outcome of the project indicates that if the development of a method to relate discharge quality to biological is to be pursued, a specifically targeted sampling approach will probably be required.

KEY WORDS

Effluent quality, biological quality, multivariate techniques.

1. INTRODUCTION

This document is the Final Report from Regional Operational Investigation 579, to quantify the relationship between effluent quality and the biological quality of rivers. This investigation was initiated to address the need to develop a methodology for relating sewage treatment works (STW) effluent quality to biologically assessed water quality.

Procedures already exist within Anglian Region for identifying STW effluents with the greatest potential impact in terms of chemical quality. This takes the form of the Index of Discharge Impact (IDI), which is calculated from the statistics of compliance with the River Needs Consent (RNC) and from an assessment of compliance of receiving waters with quality standards. The RNC is a working estimate of the consent that may be needed in future to achieve Water Quality Objectives. The IDI is used to prioritise discharges for targeting for improvement and the ability to link biological data in with the assessment would substantially enhance confidence in the methodology, hence the inception of this project.

The objectives of the work were:

Stage 1:

1. To collate existing details and data from sources within Anglian Region and construct databases;

2. To assess the factors influencing biological quality and determine the relationships between effluent quality and biological quality;

Stage 2:

3. To develop a protocol to assess the impact of STW discharges, incorporating biological data;

4. To validate the protocol methodology using data from selected sites.

Stage 3:

5. To produce a final report for the project, incorporating the findings from Stages 1 and 2.

It was recognised from the inception of this project that a successful outcome was dependent on it being possible to establish a sufficiently well defined relationship between effluent quality and biological quality to allow the former to be used as a predictor of the latter. Without such a relationship, it would not be possible to formulate the assessment protocol envisaged under Stage 2 above.

On initial completion of Stage 1 of the project, the database created was not of sufficient size, nor was there sufficient range of STW impacts, to ascertain what, if any, relationship exists between effluent quality and biological quality. Consequently it was agreed to increase the database from Anglian Region data, in the hope that this would allow the identification and

quantification of the relationship. Unfortunately, it was still not possible to adequately define this relationship, and because of this the project was terminated.

This report documents the development of the database, the statistical approaches used for the analysis of the dataset, and the reasons for the failure to find a relationship. The account of statistical analyses is for the second, augmented dataset. An account of the analysis of the original, smaller dataset can be found in the project Interim Report (Ref. CO4093), although note that the outcome was effectively the same.

2. THE DATASET

2.1 Data received

Data requirements were identified and agreed, and the data was supplied to WRc in electronic format. The supplied data derived from a number of Anglian Region databases and comprised the following:

- Details of Anglian Water Services (AWS) STW discharges: consents, u/s and d/s chemical sampling points and gauging stations, total population equivalent (TPE), dry weather flow (DWF).
- Effluent quality data
- GQA class data
- RQO data
- Biological sample results data
- Biological sample point details
- Biology species codes
- RIVPACS predictions and classifications
- River Needs Consents flows
- Index of Discharge Impact
- GQA mean and standard deviation, confidence of class and confidence of an up/down grade.
- Mean river flow (MRF)

2.2 Manipulation of data and construction of database

The primary aims of the database design and construction were to allow easy identification of those STW effluents that had associated biological monitoring sites both upstream and downstream and to produce output files containing relevant effluent and river data for statistical analysis.

Microsoft Access was used for the construction of the database, as this package could cope with the range of data formats that were supplied to WRc by Anglian Region, and it could also produce the required output files to be used in the statistical analysis. The database construction was undertaken in a number of steps which are detailed below:

1. The raw data files were imported into Access and the relationships between tables checked. Duplicate data entries, mismatching codes etc. were checked with Anglian Region and either corrected or removed.

2. A database relating the STWs with their associated upstream and downstream biological monitoring sites was produced. This was achieved by locating all the STW and biological monitoring sites on OS maps and recording the appropriate National Grid Reference and site code of the upstream and downstream biological monitoring site against the STW site code in the Relate database.

3. On completion, the Relate database was sent to Anglian Region for auditing. At this point it became apparent that a relate database already existed in one of the Anglian Areas. It was decided that this would be a more reliable source of information and was used instead of the WRc Relate database. A request was also made at this point for river flow data for STW receiving waters to be supplied.

4. Tables were constructed for the other data types (biological predictions, GQA, flow) and links established so that all data could be related to STW code.

5. Finally, data headings were standardised (as there was considerable inconsistency in heading names depending on the year that the data was produced). Also, where appropriate, data from several years was combined into one table.

Once the database construction and checking was complete, the following data files were exported in an appropriate format for statistical analysis:

- Effluent chemistry data and mean dry weather flow linked to STW code
- Upstream biological data (excluding abundance) linked to STW code
- Downstream biological data (excluding abundance) linked to STW code
- Upstream Biological predictions linked to STW code
- Downstream Biological predictions linked to STW code
- Upstream GQA chemical data linked to STW code
- Downstream GQA chemical data linked to STW code
- Mean river flow linked to STW code.

2.3 Original dataset

In the original Relate database developed, there was a total of 328 STWs. Of these, 310 had both upstream and downstream biological monitoring sites (Table 2.1).

Table 2.1 Number of STWs in original database with associated biological monitoring sites

Sites associated with STWs	Number of STWs
Total number of STWs	328
Downstream biological monitoring site	325
Upstream biological monitoring site	312
Upstream and downstream monitoring sites	310

Although there were 310 sites with both upstream and downstream biological monitoring sites, many of the upstream monitoring sites had no associated data in the biological data tables. Table 2.2 shows the number of STWs with associated biological and other data.

Table 2.2 Number of STWs in original database with data at associated upstream and downstream sites

Data type	Number STWs with Upstream data	Number of STWs with Downstream data
Biology	176	325
Biological predictions	116	148
GQA river chemistry	59	79
Mean river flow	259	259

Table 2.3 shows the progressive reduction of the size of the available dataset as different data components were introduced. The final number of STWs with all available data was 29, although it should be noted that for some of these sites some of the data was limited.

Table 2.3 Number of STWs in original database with all associated data

Data types associated with STWs	Number of STWs
Effluent, MRF, DWF	204
Effluent, MRF, DWF, Biology up and downstream	126
Effluent, MRF, DWF, Biology, RIVPACS up and downstream	75
Effluent, MRF, DWF, Biology, RIVPACS, GQA upstream	32
Effluent, MRF, DWF, Biology, RIVPACS, GQA up and downstream	29

2.4 Augmented dataset

Following the failure to find any significant relationships between effluent quality and biological quality, Anglian Region undertook further checking of the database to identify additional sites, or missing data sets, with which to augment the dataset. The augmented dataset contained 78 STWs. When the sampling data were aggregated and matched by STW, year of sampling and season of sampling, the number of STWs fell to 53. These 53 sites were used in the statistical analyses that are reported in the following sections. Note that when season and year were not used to match the data, and the analysis was performed on all 78 sites the same patterns were seen in the plots and the results of the analyses were very similar.

For one site, river flow data was not received. An estimate of flow was made for this site and it was included in the analysis. It appears in an outlier position on the plots, as a site with a large downstream decrease in biological quality. However, as the estimated effluent impact on river quality was not high, in most cases its position does not exert a strong influence on the relationships.

3. ASSESSMENT OF IMPORTANT FACTORS

3.1 Introduction

Following the auditing and revision of the database (See Section 2.2), a detailed statistical analysis was undertaken of the data from the 53 STWs for which the full set of determinands was available. The first part of the analysis involved assessing the important factors, prior to attempting to identify and quantify relationships. The determinands available for the analysis were:

Biological determinands:

- BMWP score upstream and downstream of the STW,
- ASPT upstream and downstream, and
- Lincoln Quality Index (LQI) upstream and downstream.
- RIVPACS predicted BMWP score upstream and downstream,
- RIVPACS predicted ASPT upstream and downstream.

Effluent determinands:

- total population equivalent of the STW,
- dry weather flow of the STW,
- effluent BOD concentrations (mg l^{-1}),
- effluent ammonia concentrations (mg l^{-1}),
- effluent suspended solids concentrations (mg l^{-1}).

River determinands:

- GQA BOD concentrations upstream and downstream of the STW,
- GQA ammonia concentrations upstream and downstream,
- GQA dissolved oxygen concentrations upstream and downstream,
- Mean river flow.

3.2 Biological determinands

One possible problem identified in the preliminary analysis of the initial database was that the variability in biological scores could be accentuated by the spatial differences between biological sample sites. In particular, upstream-downstream differences could be due in part to differences in sampling site characteristics. To allow for this spatial variability RIVPACS predictions for BMWP and ASPT were used as indications of the scores which the sites would be expected to attain in the absence of anthropogenic impact. Predictions were not available for the LQI scores.

In the original analysis, two slightly different ways of combining the observed and the predicted biological scores were used in the statistical analyses:

(a) the ratio of the observed to expected scores (BMWP/RIVPACS and ASPT/RIVPACS),

(b) the difference between observed and predicted (BMWP-RIVPACS and ASPT-RIVPACS).

For the second analysis, reported here, the former of these two measure was used, because it is already used by the Agency as an Ecological Quality Index (EQI), and both methods gave the same outcome.

To compare the biological quality of the downstream sites with those of the upstream, the ratio of downstream to upstream EQI was used (e.g. BMWP EQI downstream / BMWP EQI upstream). For LQI, the difference between downstream and upstream sites was used in all cases.

Although RIVPACS predictions were used to reduce variations in biology that may be related to differences in site characteristics, there is the potential danger that if the error associated with the predictions is large this may obscure any relationships between biology and effluent quality. As a check on this possibility during the original analysis, the relationships between biology and both theoretical increases in river chemistry and changes in GQA means were investigated, using the downstream-upstream differences in observed biological scores alone (e.g. downstream BMWP - upstream BMWP). This showed no apparent improvement in the relationships between biology and chemistry, so for the repeat analysis reported here the RIVPACS predictions were used.

3.3 Effluent chemistry determinands

We are interested in the deterioration of river water quality resulting from STW effluent discharges, and so the aim is to construct some measure of the potential effect that the STW has on the chemistry of the receiving water. Knowing the concentrations of the effluent alone does not give information about how much the water quality has decreased as a result of the discharge. If we have a mean chemical concentration in the effluent of x mg l^{-1} and a flow of DWF 1000m^3/day then there should be approximately x × DWF × 1000 kg/day of the chemical entering the receiving water. If the river has a mean concentration of y mg l^{-1} and a flow of MRF 1000m^3/day then there should be approximately y × MRF × 1000 kg/day of the chemical

flowing past the discharge point. The concentration in the receiving water immediately downstream of the discharge should therefore be approximated by,

$$\frac{x \times DWF + y \times MRF}{DWF + MRF} \text{ mg } l^{-1}.$$

Therefore the theoretical mean increase in the receiving water concentration is

$$\frac{(x - y) \times DWF}{DWF + MRF} \text{ mg } l^{-1}.$$

Mean river flow data (MRF, the flows in the receiving waters) and dry weather flow data (DWF, the flows from the discharges) were available for many of the STWs in the database. In order to calculate the theoretical increase in receiving water concentrations, measures of the chemical concentrations in the receiving waters upstream of the discharges are needed. To this effect, mean BOD and ammonia concentrations calculated for GQA purposes were used, thus allowing average theoretical increases to be estimated for BOD and ammonia. Unfortunately there was no suitable information for suspended solids and so the increase was estimated by

$$\frac{x \times DWF}{DWF + MRF} \text{ mg } l^{-1}$$

which is the maximum increase possible (i.e. if there were no suspended solids at all in the receiving water upstream of the discharge).

3.4 Examination of possible relationships

3.4.1 EQI and theoretical increase in river chemistry

Figures 3.1 - 3.4 show the ratios of downstream to upstream EQIs (BMWP and ASPT) plotted against the theoretical increases in BOD and ammonia concentrations in the receiving waters. Note that there are a number of data points for each STW. There is no evidence of any clear trend between biological and chemical variables. Most of the data points fall in a tight cluster in the centre of the plots, representing STWs with little impact on receiving water quality and little change in biological quality. The hope was that the dataset would include STWs having a wider range of impacts, and particularly more STWs having a large impact on receiving water chemistry. However, the few sites in these plots where the theoretical increases in BOD and ammonia are high, exhibit little change in biological quality or, if anything, improved biological quality downstream. Moreover, for the STWs showing little, or no theoretical impact on river quality, there is considerable spread in the EQI measure, over more than an order of magnitude, indicating that it is inherently variable.

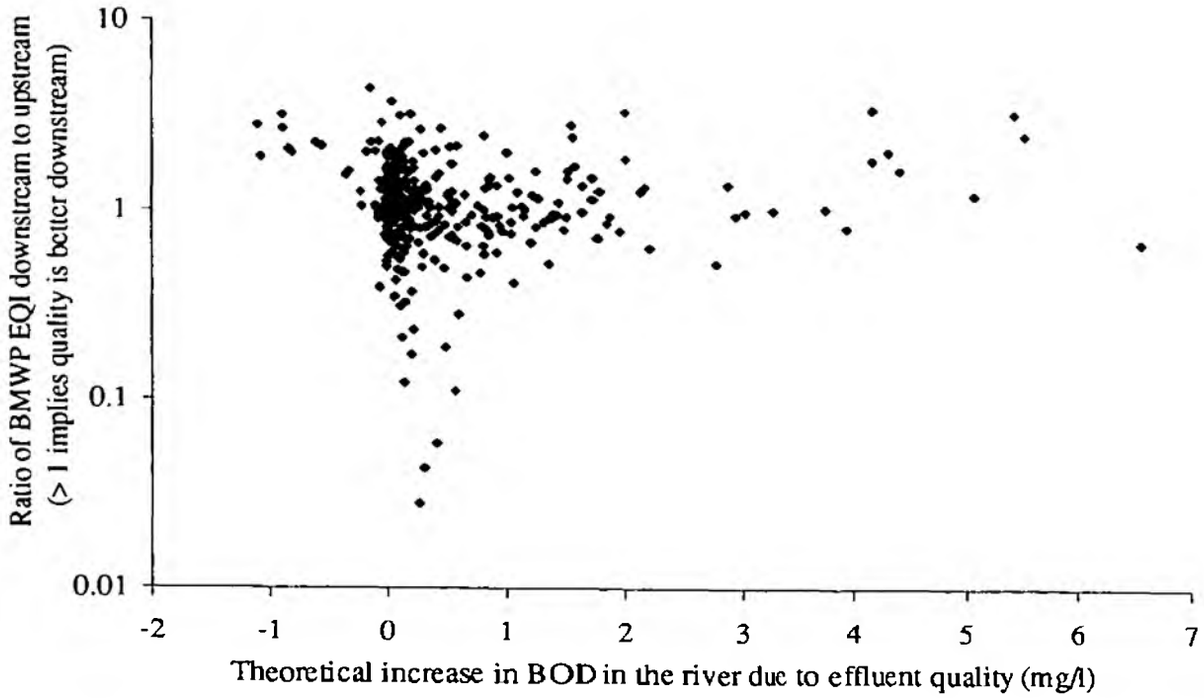

Figure 3.1 Relationship between BMWP EQI ratio and theoretical BOD increase

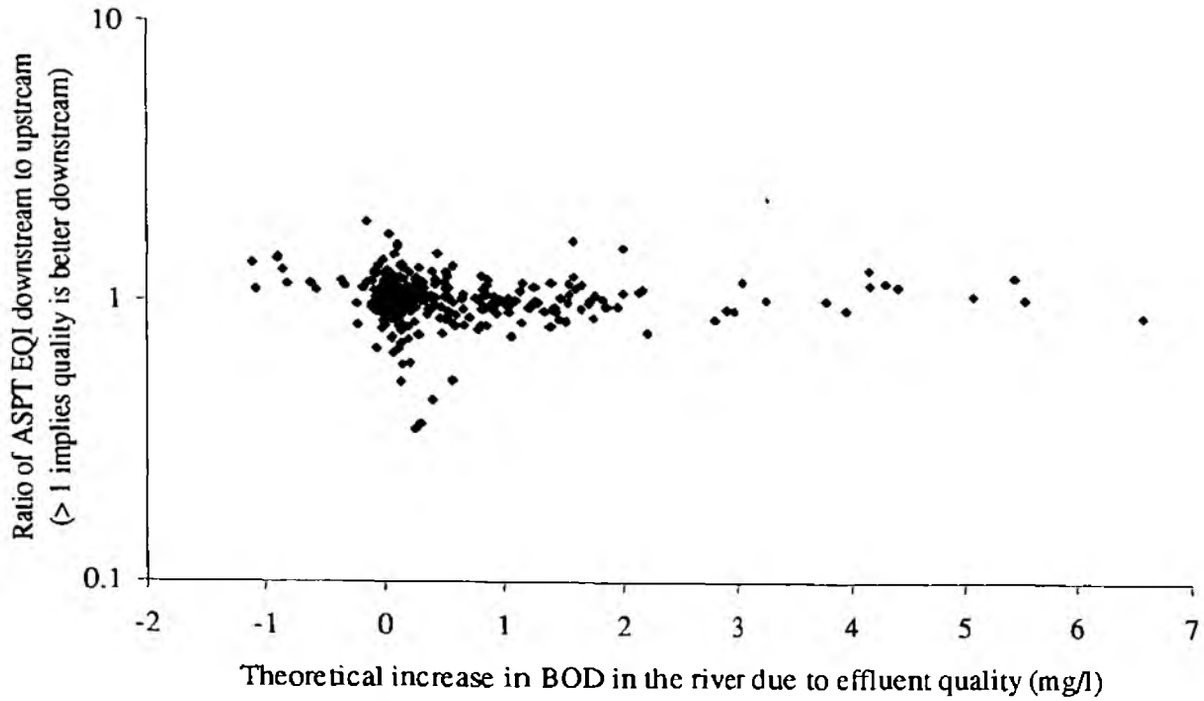

Figure 3.2 Relationship between ASPT EQI ratio and theoretical BOD increase

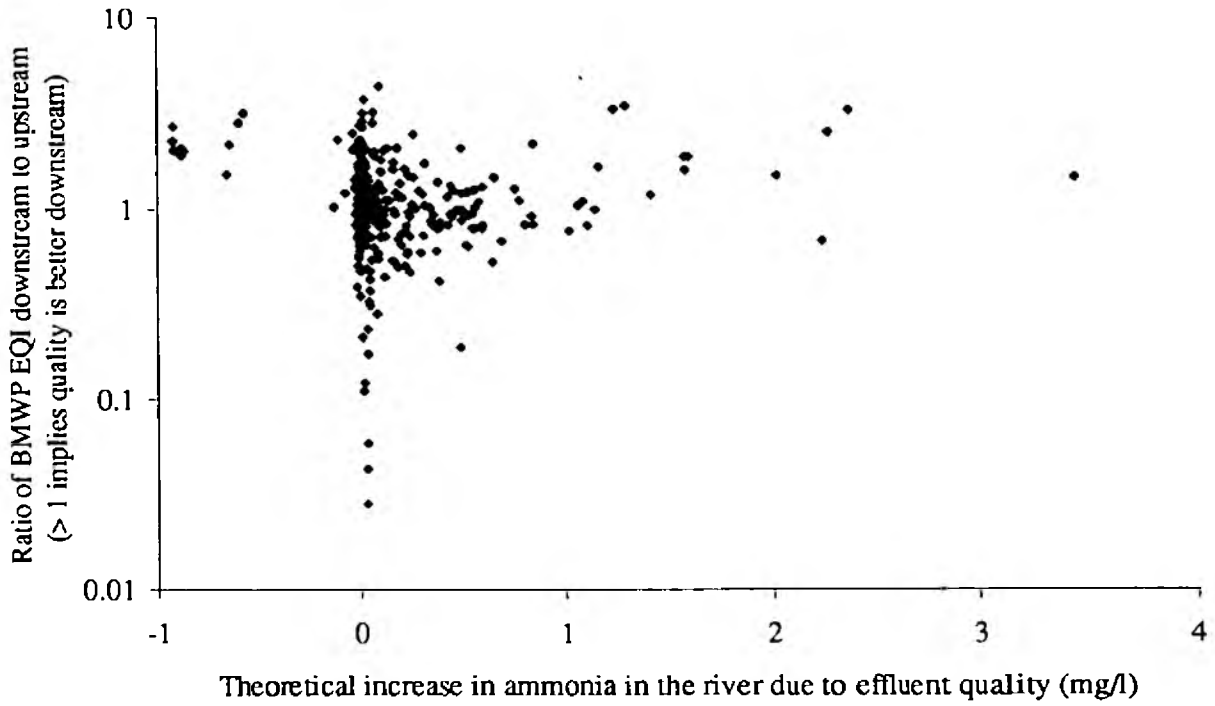

Figure 3.3 Relationship between BMWP EQI and theoretical ammonia increase

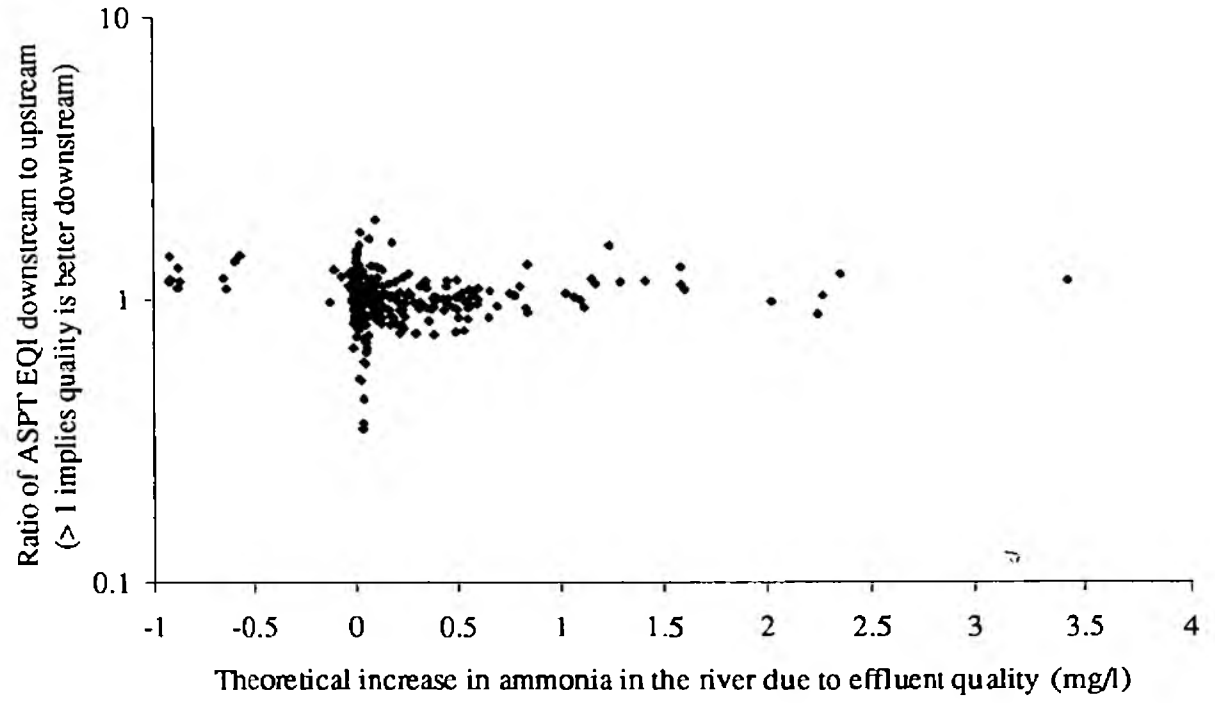

Figure 3.4 Relationship between ASPT EQI and theoretical ammonia increase

3.4.2 EQI and change in GQA means

In Figures 3.5 - 3.10 the ratios of EQIs are plotted against the increases in the chemical concentrations in the receiving water going from upstream to downstream. These chemical increases are estimated from the GQA means for BOD, ammonia and dissolved oxygen. One would expect any relationships between chemical quality and biology to be revealed in these figures. However, any relationships found are weak, for two reasons. Firstly, for BOD, and especially for ammonia, the data points cluster in the area where there is little change in chemistry and any relationship is influenced by relatively few outlying points. For dissolved oxygen there is more spread but, here, the second reason is apparent, namely the inherent variability of the EQI ratios, particularly for BMWP.

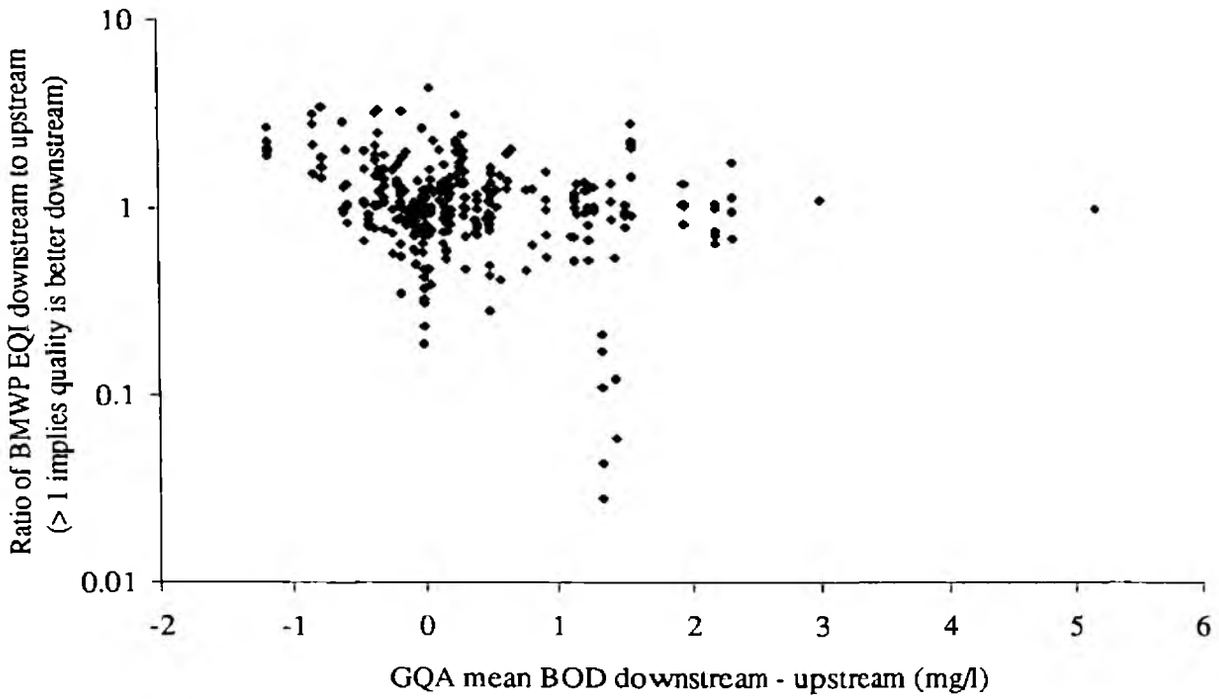

Figure 3.5 Relationship between BMWP EQI ratio and change in GQA BOD

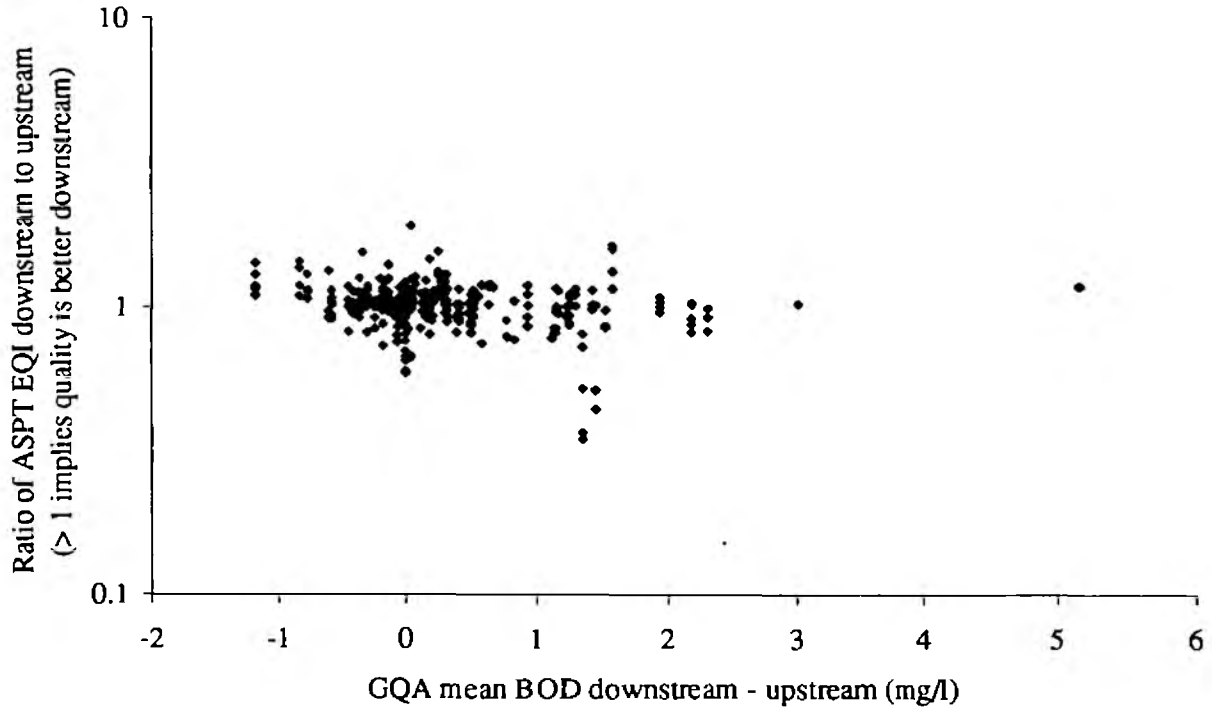

Figure 3.6 Relationship between ASPT EQI ratio and change in GQA BOD

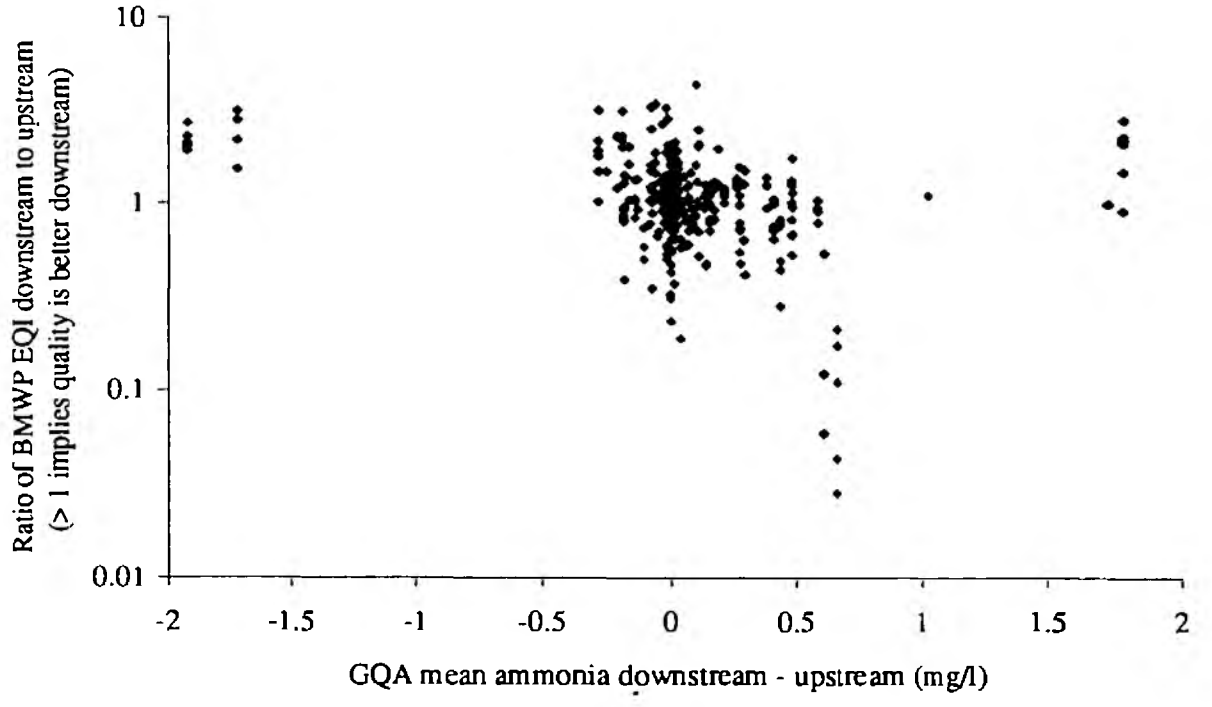

Figure 3.7 Relationship between BMWP EQI ration and change in GQA ammonia

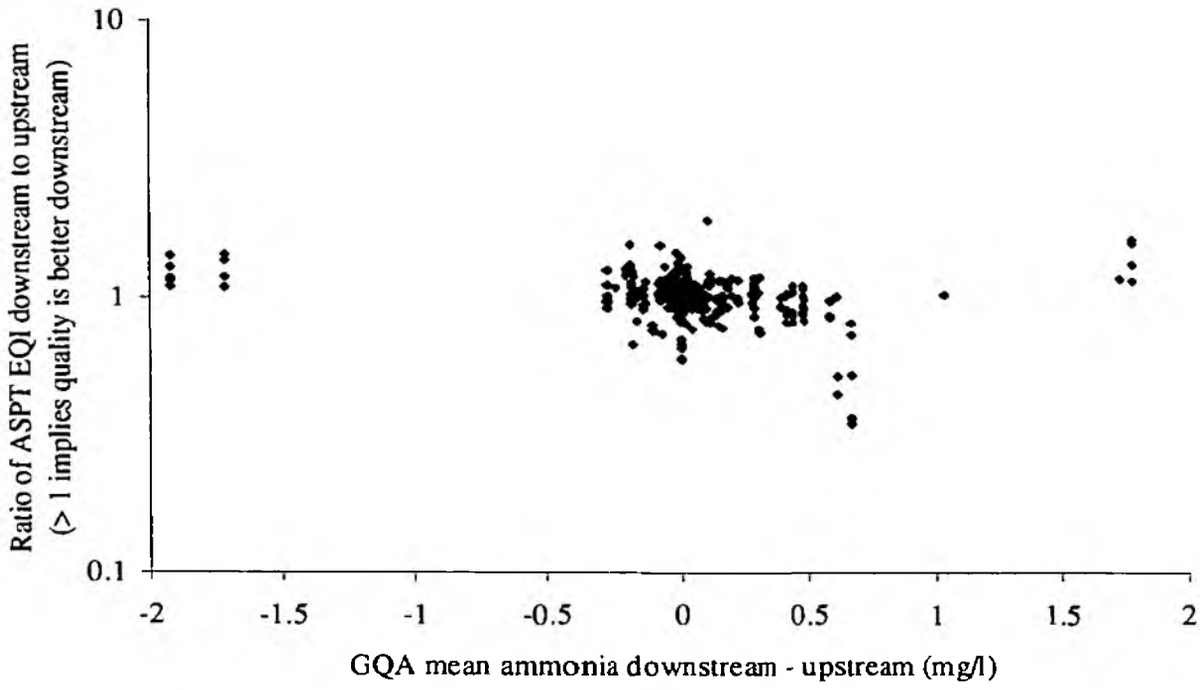

Figure 3.8 Relationship between ASPT EQI ratio and change in GQA ammonia

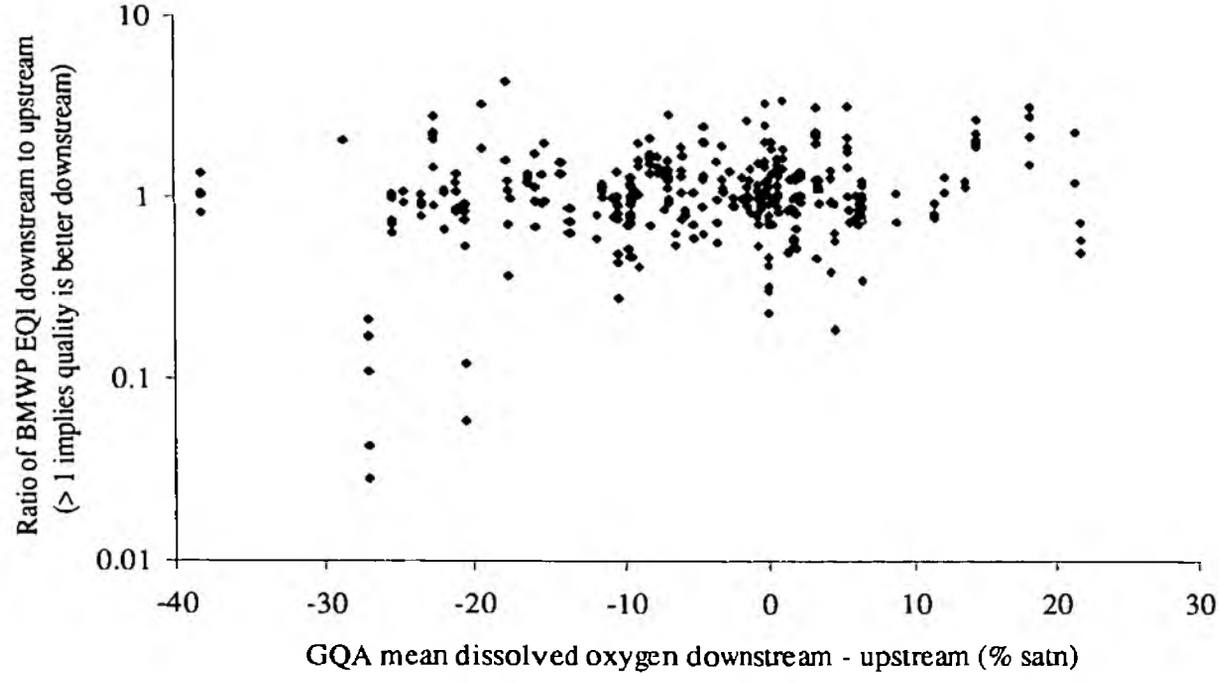

Figure 3.9 Relationship between BMWP EQI ratio and change in GQA dissolved oxygen

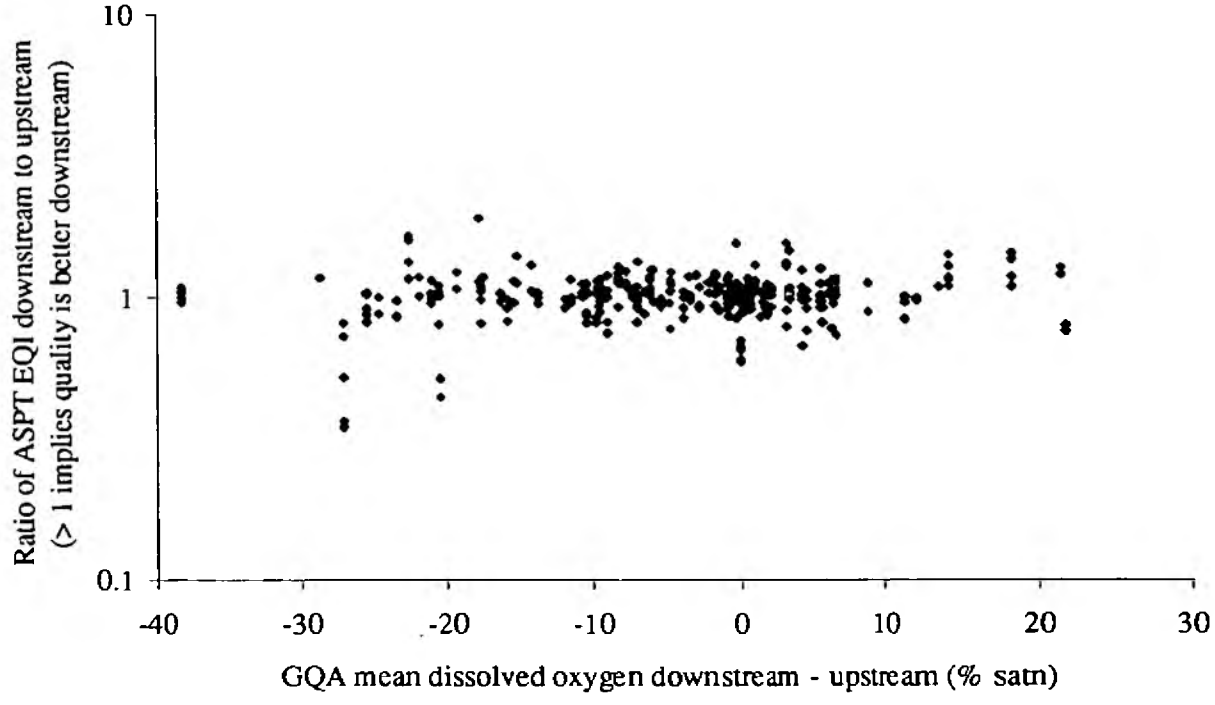

Figure 3.10 Relationship between ASPT EQI ratio and change in GQA dissolved oxygen

3.4.3 GQA mean and theoretical increase in river chemistry

Figures 3.11 and 3.12 show plots of the changes in GQA means for BOD and ammonia against the theoretical changes in river chemistry due to the effluents. There are no obvious relationships between change in quality as measured by GQA and that predicted from effluent quality. Again, the majority of the data points are in the 'no impact' cluster, although there are a few data points for STWs with higher theoretical increases of BOD or ammonia in the receiving water. However, these generally do not correspond with actual differences according to the GQA data.

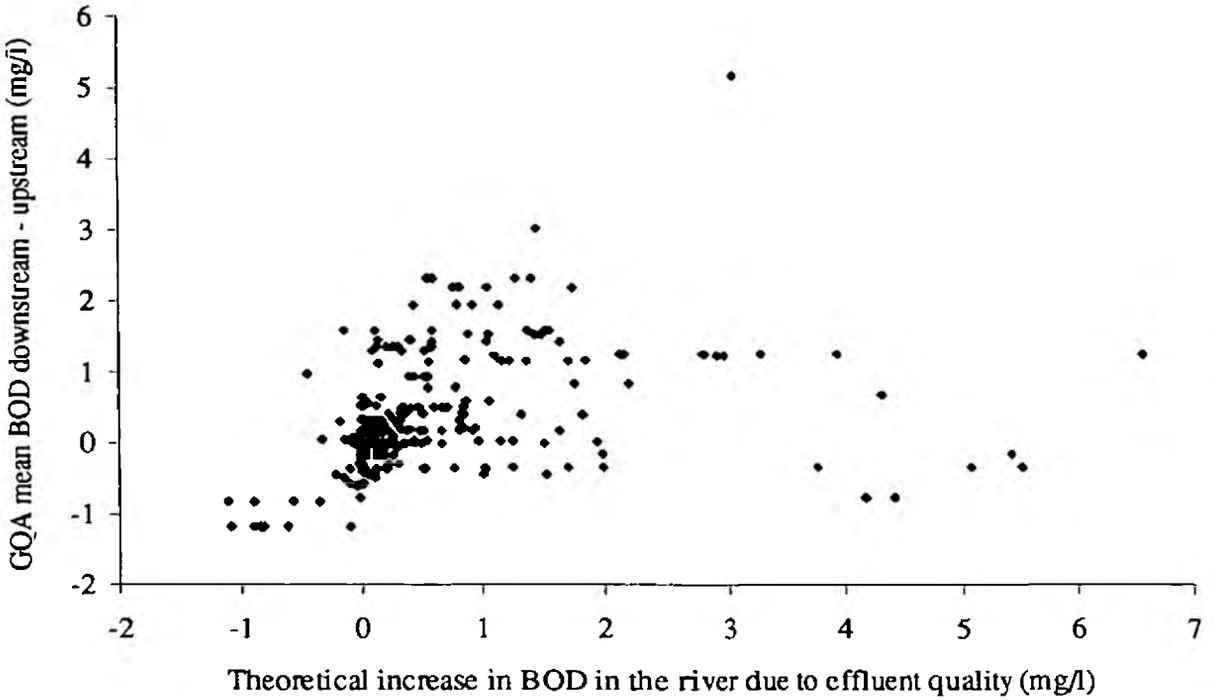

Figure 3.11 Relationship between GQA change and theoretical increase for BOD

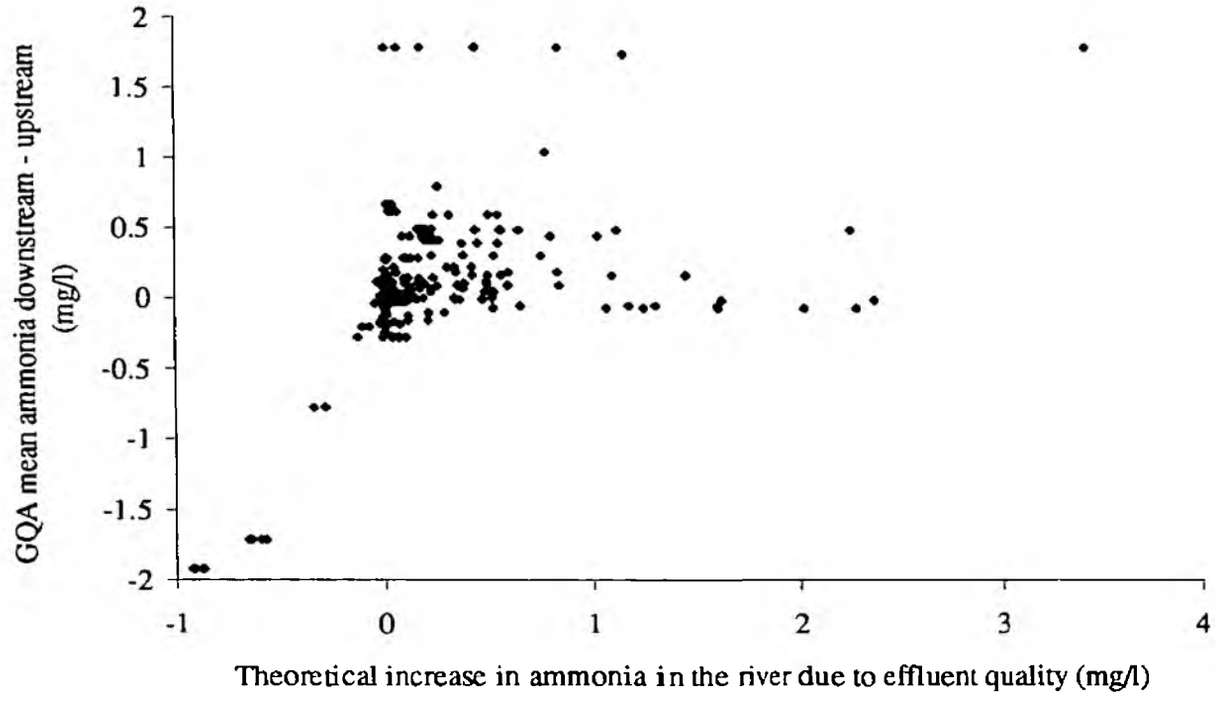

Figure 3.12 Relationship between GQA change and theoretical increase for ammonia

3.4.4 Changes over time

In Figure 3.13 the theoretical increase in BOD due to the effluent discharge is plotted against year. The connecting lines join samples from the same STW. With the exception of a few STWs most of these theoretical increases remain fairly constant through time. This pattern is also seen in Figure 3.14 where the differences in GQA mean BOD are plotted. The plot of the ratio of ASPT EQIs against time in Figure 3.15 shows a similar constancy. From these diagrams there do not appear to be any significant overall trends in either the effluent quality, the GQA means or biology. The same pattern is found when looking at the other biological and chemical measures.

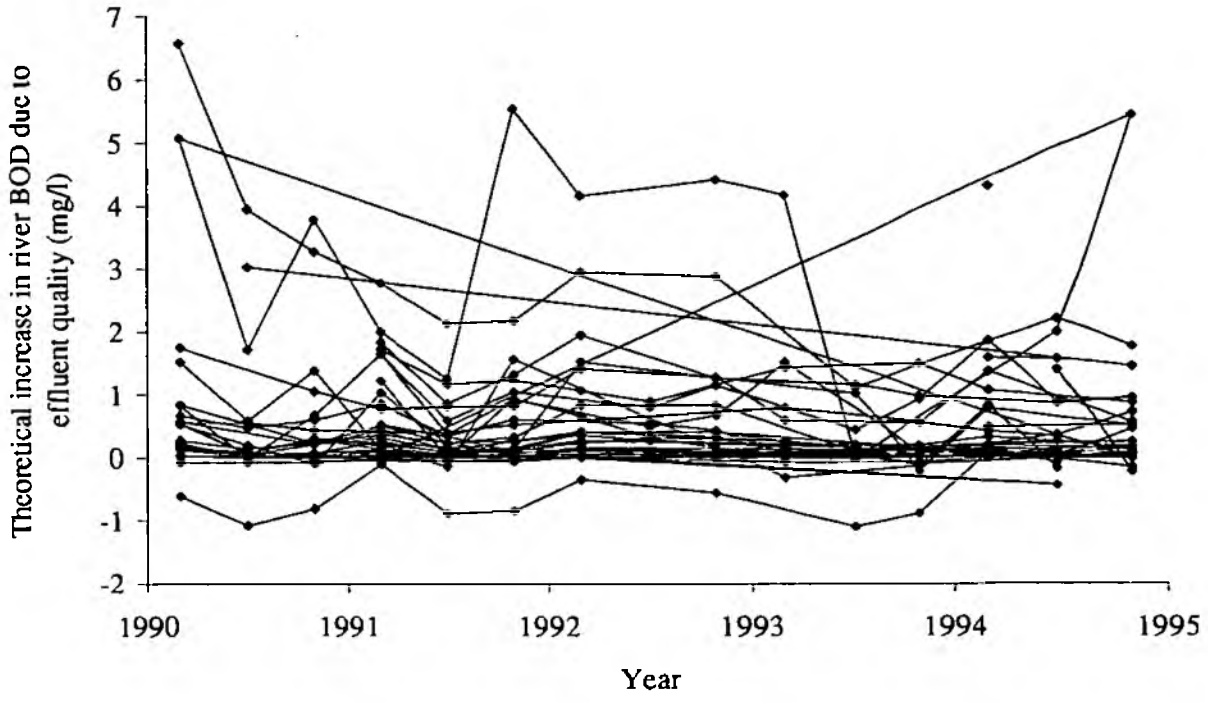

Figure 3.13 Theoretical increase in BOD plotted against time

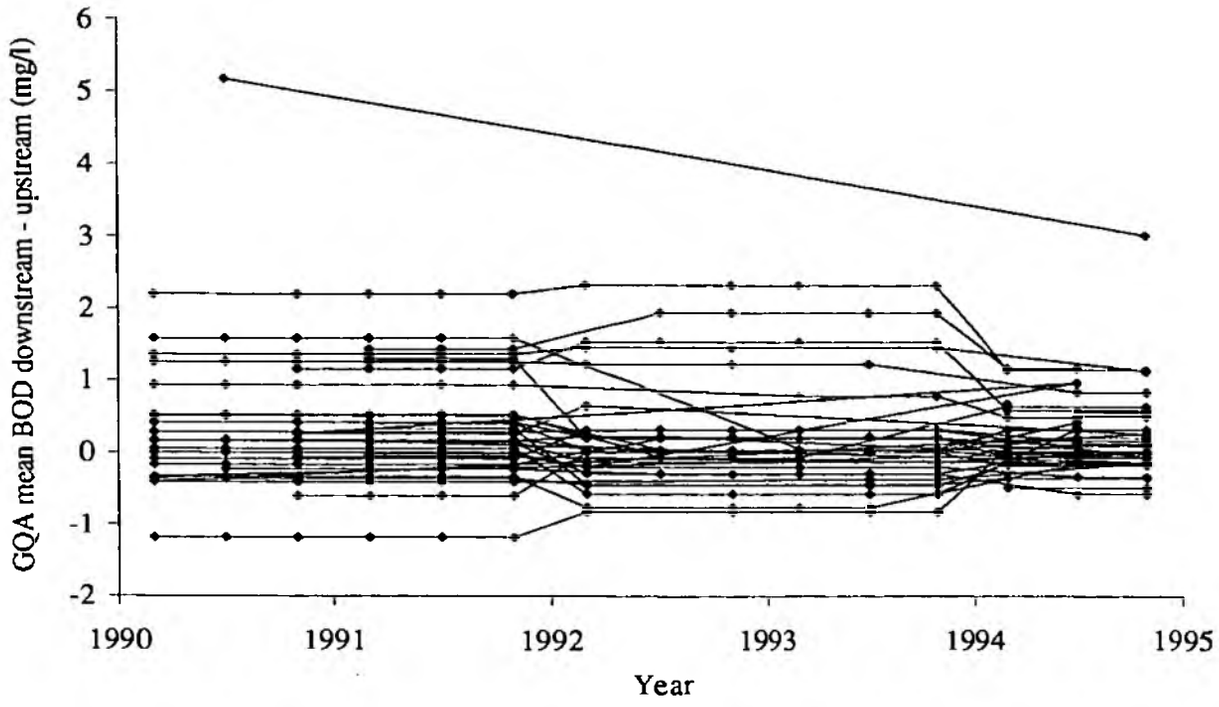

Figure 3.14 GQA BOD change plotted against time

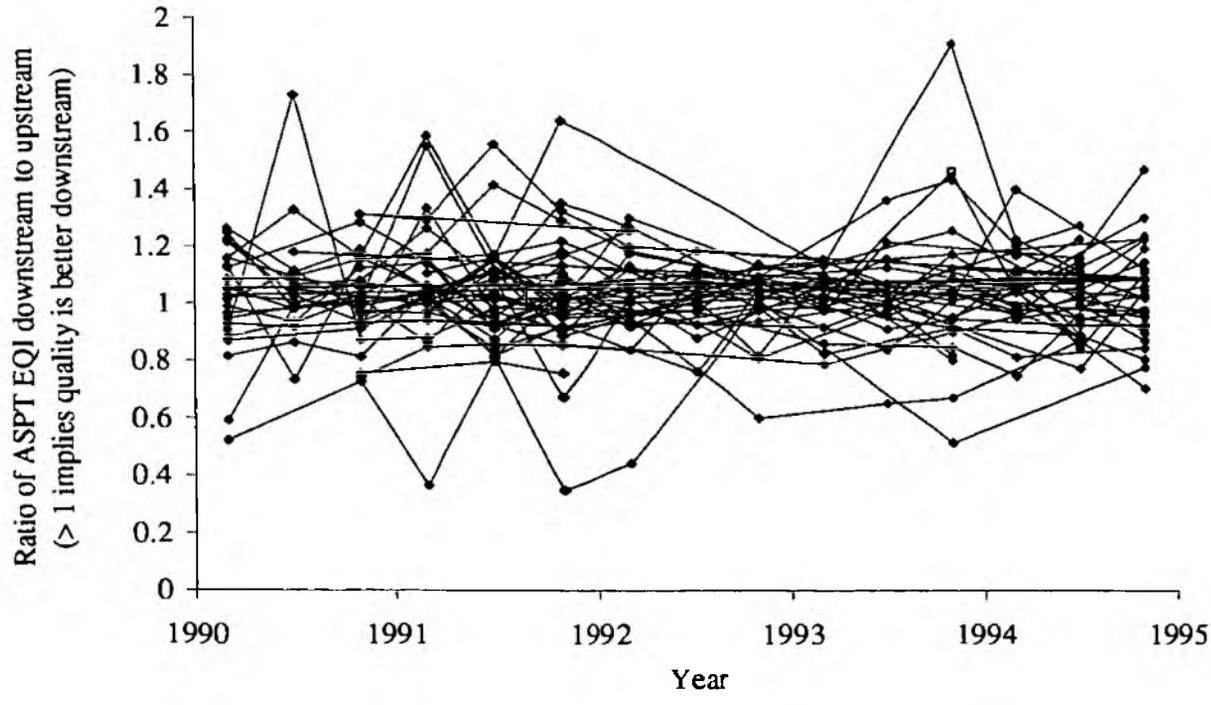

Figure 3.15 Ratio of ASPT EQI plotted against time

4. IDENTIFICATION OF RELATIONSHIPS BETWEEN EFFLUENT QUALITY AND BIOLOGICAL QUALITY

4.1 Introduction

Interest lies in the relationship between the effluent quality of discharges and the changes in biological quality downstream of discharges. In order to define this relationship optimally, it is necessary to remove as many sources of variability which are not part of the relationship as is possible. To achieve this, other relevant factors and determinands were introduced into the analysis. Some of these have already been used in constructing the 'biological changes' and the 'theoretical increases in receiving water concentrations' factors. In addition to these, a range of other factors were used to account for spatial and temporal variability.

The explanatory variables which were used for this part of the analysis were:

- theoretical increases to receiving water concentrations (BOD, ammonia and suspended solids),
- season,
- year,
- STW (as a block effect),
- GQA means for upstream sites (BOD, ammonia and dissolved oxygen),
- RIVPACS predicted biology up and downstream (BMWP and ASPT), and
- observed biology upstream (BMWP, ASPT and LQI).

The change in biology variables used were:

- Log (BMWP EQI downstream / BMWP EQI upstream) where BMWP EQI is observed BMWP / predicted BMWP,
- Log (ASPT EQI downstream / ASPT EQI upstream),
- LQI downstream - LQI upstream

Season and year are included to try to account for any possible temporal variability. STW is included to account for some of the spatial variability. The GQA means upstream, the upstream biology and RIVPACS predictions are included, since a river which is already polluted may not show as large a deterioration in biological quality as one which is relatively clean.

The first step in this analysis is to reduce the problem to a univariate regression problem (although still a multiple regression) to determine, where possible, relationships might lie and

which variables appear to have the most influence. This reduction of the problem essentially throws away some of the information in the data in order to arrive at the simpler form. This information loss is minimised, however, by using multivariate techniques (principal coordinates analysis - a form of metric scaling).

The second part of the analysis attempts to use all of the information in the data by using a canonical correlation analysis to link the biological-change variables to the set of explanatory variables.

4.2 Principal coordinates analysis

4.2.1 Introduction

Principal coordinates analysis is commonly used for reducing the dimensionality of data sets where there are too many variables, or where the intention is to describe the data set with a single measure or pair of measures. For the data considered here, there are three variables describing biological changes (the ratio of EQIs for BMWP and ASPT and difference between downstream and upstream LQI) which are to be related to the changes in river water quality. If there was only one variable describing biology then it would be straightforward to perform some sort of regression analysis to estimate the relationship. A principal coordinates analysis applied to these biological variables would provide three transformations (weighted averages) of the originals. The first of these transformed variables would contain as much of the variation from the original three as is possible using a linear transformation. The second transformed variable would contain as much of the variation left over from the first as was possible, and the last would contain the remainder of the variation. If the first transformed variable (called the first principal coordinate) contains the majority of the variation contained in the data set then it means that the three variables can be reduced to a single one without losing very much information. In other words, the original three variables were all telling the same story and so nothing was to be gained by having all three (although all three are still there as part of the first principal coordinate). See Mardia *et al.* (1979) for a more detailed explanation.

4.2.2 Principal coordinates analysis of the biological variables

The intention here is to reduce the Biological variables to a single measure and regress this single variable on the explanatory variables. The Biological variables are the standardised Log-transformed EQI ratios (downstream over upstream EQI) for BMWP and ASPT, and the standardised differences between downstream and upstream scores for LQI. Standardisation is achieved by subtraction of the mean and division by the standard deviation.

Table 4.1 shows how much of the information contained in the three Biological variables can be explained by each of the principal coordinates. The first principal coordinate explains nearly 85% of the variability of the Biological variables (i.e. it contains 85% of the information), and the first and second coordinates together explain nearly 95% of the variability. This means that we could replace these three Biological variables by their first one or two principal coordinates and lose very little of the information they contain.

Table 4.1 Percentage variation explained by the principal coordinates of the biological changes variables

	Principal coordinate		
	1st	2nd	3rd
% of variation explained	84.5%	9.9%	5.6%

4.2.3 Principal Coordinates Analysis of the Effluent Variables

The Effluent variables are the standardised theoretical increase in BOD and ammonia, and the standardised maximum theoretical increase in suspended solids. Table 4.2 shows how much of the information contained in the Effluent variables can be explained by each of the principal coordinates.

Table 4.2 Percentages of the total variation explained by the principal coordinates of the effluent variables.

	Principal coordinate		
	1st	2nd	3rd
% of variation explained	78.0%	15.6%	6.4%

The first principal coordinate explains 78% of the variability of the Effluent variables (i.e. it contains 78% of the information), whereas the first and second coordinates explain nearly 94% of the variability between them. The first two principal coordinates, therefore, can be used to represent the Effluent variables without much loss of information.

4.2.4 Principal Coordinates Analysis of the Upstream GQA Variables

The Upstream GQA variables are the standardised GQA means of upstream Log-BOD, Log-Ammonia, and Dissolved Oxygen. Table 4.3 shows the amount of variation contained within the Upstream GQA variables that is explained by each of the principal coordinates.

Table 4.3 Percentages of the total variation explained by the principal coordinates of the upstream GQA variables.

	Principal coordinate		
	1st	2nd	3rd
% of variation explained	61.5%	28.0%	10.5%

The first principal coordinate explains nearly 62% of the variability of the Upstream GQA variables, whereas the first and second coordinates explain nearly 90% of the variability between them. The first two principal coordinates, therefore, can be used to represent the upstream GQA variables without much loss of information.

4.2.5 Principal Coordinates Analysis of the Upstream Biology Variables

The Upstream Biology variables are the standardised upstream BMWP scores and ASPT, and the standardised LQI scores. Table 4.4 shows how much of the information contained in the upstream biology variables can be explained by each of the principal coordinates.

Table 4.4 Percentages of the total variation explained by the principal coordinates of the upstream biology variables.

	Principal coordinate		
	1st	2nd	3rd
% of variation explained	86.6%	7.8%	5.7%

The first principal coordinate explains 87% of the variability of the Upstream Biology variables and so it can be used to represent this variable set without a large loss of information.

4.2.6 Principal Coordinates Analysis of the Upstream RIVPACS Prediction Variables

The Upstream RIVPACS Prediction variables are the standardised upstream RIVPACS-predicted BMWP scores and ASPT. Table 4.5 shows the amount of information within the Upstream RIVPACS variables that can be explained by each of the principal coordinates.

Table 4.5 Percentages of the total variation explained by the principal coordinates of the upstream RIVPACS predictions variables.

	Principal coordinate	
	1st	2nd
% of variation explained	86.0%	14.0%

The first principal coordinate explains 86% of the total variation and so can be used to represent the Upstream RIVPACS Prediction variables without a large loss of information.

4.2.7 Regression of Biological 1st Principal Coordinate on explanatory variables including upstream biology scores and STW effect

The first principal coordinate (PC) of the Biological variables was regressed on various explanatory variables. These variables were the first two principal coordinates of the Effluent variables, the first two principal coordinates of the Upstream GQA variables (mean of Log BOD, Log ammonia and Log suspended solids), the first principal coordinate of the Upstream Biology variables, the first principal coordinate of the Upstream RIVPACS prediction variables, and factors representing season, year and STW. A stepwise backwards selection procedure was used to eliminate explanatory variables whose effects were not significant. The resulting model is given below.

Response variate: 1st PC of standardised Log-EQI ratio.

Explanatory variables: Constant, 1st PC of upstream GQA, 1st PC of upstream biology, year, and site.

Table 4.6 shows the summary analysis of variance table for the regression.

Table 4.6 Analysis of variance table for the regression

Source	d.f.	Sums of squares	Mean squared	F-ratio	P-value
Regression	52	619.1	11.91	15.23	<.001
Residual	269	210.3	0.78		
Total	321	829.4	2.58		

Percentage variance accounted for: 69.7%
Standard error of observations is estimated to be 0.884

Table 4.7 shows the estimates of regression coefficients for the explanatory variables, and Table 4.8 shows the accumulated analysis of variance from the stepwise selection procedure.

Table 4.7 Estimates of regression coefficients (STW effect levels are not shown)

Explanatory variable	Estimate	Standard error of estimate	t statistic (d.f. = 268)
Constant	-1.18	0.908	-1.30
1st PC of upstream GQA	0.34	0.140	2.39
1st PC of upstream biology	-0.65	0.055	-11.79
Year effect - 1991	-0.10	0.143	-0.67
Year effect - 1992	0		
Year effect - 1993	-0.35	0.172	-2.04
Year effect - 1994	-1.01	0.236	-4.29

Table 4.8 Accumulated analysis of variance: contributions to the sums of squares from adding each explanatory variable to the model in the order listed (+ sign), and then dropping them in the order listed (- sign).

Change	d.f.	Sum of squares	Mean square	F-ratio	P-value
+1st PC of Effluent	1	4.55	4.550	5.80	0.017
+2nd PC of Effluent	1	35.09	35.085	44.69	<.001
+1st PC of upstream GQA	1	15.15	15.146	19.29	<.001
+2nd PC of upstream GQA	1	5.27	5.273	6.72	0.010
+1st PC of upstream Biology	1	157.19	157.19	200.24	<.001
+1st PC of upstream RIVPACS	1	10.60	10.604	13.51	<.001
+Year effect	3	23.28	7.759	9.88	<.001
+Season effect	2	2.70	1.350	1.72	0.181
+STW effect	47	369.11	7.853	10.00	<.001
Residual deviation	**263**	**206.46**	**0.785**		
-2nd PC of Effluent	-1	-0.12	0.123	0.16	0.693
-1st PC of upstream RIVPACS	-1	-0.20	0.196	0.25	0.618
-Season effect	-2	-1.37	0.685	0.87	0.419
-1st PC of Effluent	-1	-0.94	0.943	1.20	0.274
-2nd PC of upstream GQA	-1	-1.17	1.172	1.49	0.223
Total	**321**	**829.40**	**2.584**		

The effect of the effluent variables on the EQI biological variables is not significant in the framework of this regression analysis. Most of the explainable variability (only 70%) is attributable to the STW factor and upstream biology. It is not surprising that there is some correlation between upstream biology and the EQI variables since the EQIs are calculated, in part, from upstream biology. The regression analysis was repeated excluding upstream biology and upstream RIVPACS prediction. The result of this was that only the STW factor was significant and only 54.5% of the total variation was explainable. If the STW factor is excluded from the analysis as well, then the 2nd PC of the effluent variables remains in the model as a significant variate, along with the 1st PC of the upstream GQA variables. However, a mere 5.5% of the total variability is explained by this model. In other words, the predictive power of the model will be very small.

4.3 Canonical correlation

The method of canonical correlation analysis is a multivariate technique which takes a set of y-variables, in this case the biological differences, and tries to 'manipulate' them to give the best possible correlation with a set of x-variables, in this case the chemical determinands. The 'manipulation' involves combining the y-variables (using a weighted average) into a single measure, and also combining the x-variables into a corresponding single measure. These new x and y measures are called the first canonical variates. The weights (or loadings) used to make the new variables are chosen so as to maximise the correlation between the new x and y-variables. After the 1st canonical variates have been made, the method finds another set of weightings which make up the 2nd canonical variates. These weights are chosen so that the 1st and 2nd canonical variates are not correlated with each other. This process is repeated for as many canonical variates as there are variables in the smaller of the two original data sets. Each successive pair of canonical variates have smaller correlations and the hope is that most of the overall correlation (say 90%) is accounted for by the first one or two canonical variates. A more technical description of canonical correlation can be found in Mardia *et al.* (1979).

The set of y-variables used in this analysis was the biological variables set as described in Section 4.2.3, without standardisation. The set of x-variables was the full set of explanatory variables, with two exceptions. Firstly, the theoretical increases in BOD and ammonia were combined into their first principal component, since there is such a high degree of colinearity between them, and, secondly, STW cannot be included because it is a multilevel factor, which would require the inclusion of more variables than can be included in the canonical correlation.

The results of the analysis are shown in Table 4.9. The correlations shown in Table 4.9 are between the canonical variates described above. The first pair of canonical variates have a correlation of 0.72 which means that roughly 50% of the variation between the first canonical variants is explained, and the rest is 'noise'. Whilst this is an improvement on the regression of the previous section (the regression explained more but included a covariate for STW), it is still not as good a relationship as one would hope for. As the canonical correlations are all fairly similar, there is no single manipulation of the datasets which draws out a good relationship between the x and y variables. This is analogous to a plot of the data in a regression looking like a fuzzy ball, rather than the long, thin cigar shape one would see if there was a good correlation.

Table 4.9 Canonical correlations

Canonical Variable	Correlation	% of Total Correlation	Cumulative % of Correlation
1st	0.72	37.5	37.5
2nd	0.63	32.7	70.2
3rd	0.58	29.8	100.0

Table 4.10 Coefficients of variables in relationships

Explanatory variable	Log(BMWP EQI Ratio)	Log(ASPT EQI Ratio)	Difference in LQIs
1st PC of Effluent variables	-0.07	-0.02	-0.04
Effluent Maximum Theoretical Increase in Suspended Solids	0.04	0.00	0.02
GQA mean Log-BOD upstream	0.04	0.01	0.16
GQA mean Log-Ammonia upstream	-0.07	-0.02	-0.10
GQA mean Dissolved Oxygen upstream	-0.01	0.00	-0.01
RIVPACS-predicted BMWP upstream	0.01	0.00	0.00
RIVPACS-predicted ASPT upstream	0.16	0.20	0.02
RIVPACS-predicted BMWP downstream	0.00	0.00	0.01
RIVPACS-predicted ASPT downstream	0.07	-0.17	-0.37
Observed BMWP upstream	-0.01	0.00	0.01
Observed ASPT upstream	-0.10	-0.19	0.29
Observed LQI upstream	-0.05	0.00	-0.91
Summer	0.10	0.00	0.12
Winter	0.08	0.01	0.12
Year 1991	0.04	0.02	0.21
Year 1992	0.00	0.00	0.00
Year 1993	-0.11	-0.01	0.27
Year 1994	0.28	0.08	0.76
Year 1995	0.00	0.00	0.00

It is useful to examine this canonical correlation analysis in more detail to see where the explanatory power comes from. Unfortunately with canonical correlation there is no easy way to test the importance of each of the contributing variables in the analysis. However, it is unlikely that the importance of the effluent data will be any greater than in the regression. To

see the effect of changes in effluent quality as predicted by the canonical correlation, algebraic manipulation of the weightings must be done. These weights can be transformed to give, in effect, three regression equations relating the biological-change variables to the predictors. The coefficients of each of the predictor variables resulting from this manipulation can be seen in Table 4.10.

The model shows that a relatively large decrease in the effluent quality will only give a small predicted worsening in the biological quality downstream. For example, if the theoretical increases in BOD and ammonia were to be raised by 1 standard deviation each (a relatively large increase) this would result in an decrease in the Log BMWP EQI ratio of approximately 0.1. Smaller changes would be expected for LQI and ASPT.

5. DISCUSSION AND CONCLUSIONS

The Agency has a national system for biological assessment of water quality but no formal method of relating discharge quality to biologically assessed water quality. This is due to the lack of development of existing data and protocols.

The ability to assess the overall impact of a STW discharge, taking into account both chemical and biological data, would significantly improve the Agency's ability to target investment and demonstrate improvements. At present, there is no formal way of assessing discharge quality in terms of sanitary determinands and biological quality. Therefore, a need for such a methodology to assist in planned investment in sewage treatment in order to provide value for money was established, leading to the inception of this project.

In order to develop such a methodology, it would first be necessary to identify and quantify relationships between biological quality and the chemical quality of effluents and receiving waters. However, despite using a dataset representing a relatively large number of STWs, it has not proved possible to identify convincing relationships between effluent quality and biological quality in receiving waters. The primary reasons are that there was a lot of noise in the variables examined, and relationships found between variables were weak. Specifically:

- few of the STWs were actually having much impact on river chemistry, nor, by implication, on biological quality;
- biological data tended to show high variability within sites;
- other data may also have been variable, or not fully appropriate (e.g. GQA and flow monitoring sites may not be ideally sited with reference to the STWs).

Since there were only three or four STWs having a relatively high impact on river quality, their individual differences were highly influential to the outcome of any analysis. (The results can flip one way or the other by the inclusion or exclusion of one or two STWs.). Furthermore, the sites with high theoretical impact appear to show no change, or even an improvement, in biological quality downstream. There are a number of possible explanations for this, for example:

- the upstream site may be affected by some other impact;
- times of travel and timing of sampling may mean that the GQA data is biased, which in turn biases the estimate of impact;
- flow data may be inappropriate (for STWs these are DWF, for rivers they are mean river flow at a gauging station not necessarily near to the site)

If the data used in the analysis is truly representative of Anglian rivers and STWs it does suggest that most STWs have little or no effect on biological quality. However, it is difficult to exclude the possibility that a relationship does exist but is masked by the variability and imprecision inherent in much of the data used in the analyses. Moreover, the existing sampling

network from which the data derive is not generally targeted for assessing the impact of specific STWs. Although it was recognised by the Agency that it may be necessary to take an experimental approach and carry out specifically targeted monitoring to achieve a robust data set, it was felt that existing data may be sufficient and should be assessed in the first instance to avoid unnecessary effort. Although a larger set of data, covering a wider geographical area and incorporating STWs with a greater range of impacts, might allow the identification of relationships between effluent and biological quality, the indication is that a specifically targeted experimental approach would be required.

It was recognised from the outset of this project that if clear relationships could not be established between effluent quality and biological quality then there was a risk that the use of biological assessment may be discredited. The fact that this project was not able to define clear relationships can be attributed to weaknesses in the available data rather than a lack of any relationships. The use of macroinvertebrate communities in the GQA survey and in the operational assessment of water quality clearly demonstrates that a relationship does exist.

Moving away from the impact of specific effluents, there may be value in re-assessing the more general relationship between chemical and biological quality possibly using the national GQA data set. An analysis of this type using data from the 1990 quinquennial survey showed that there was such a relationship (Kinley and Gulson 1993). However, while there was a good chance of predicting chemical class to within one class from biological data, there was a poor chance of predicting the exact chemical class. Since 1990, there have been initiatives to improve the Quality Assurance of biological data, as well as an extensive rationalisation of the national sampling network. Whilst a very close relationship would not be expected, since biology provides a measure of the impact of much more than is traditionally monitored chemically, a good understanding of the relationship may be of use, for example in enhancing the ability to identify biological and chemical quality mismatches. Hence, a re-evaluation may be worthwhile, although, there is still a risk that inherent variability in data may mean that any relationships are too variable to be operationally useful.

6. RECOMMENDATIONS

It was not possible to identify appropriate relationships between effluent quality and biological quality using the Anglian Region dataset and it is recommended that this project is terminated.

To identify the relationships of interest, with a view to developing operational tools, would probably require specifically targeted sampling.

There may be value in re-assessing the more general relationships between biological and chemical quality using the national GQA data.

REFERENCES

Kinley, R.D and Gulson, J (1993) Statistical analysis of relationships between chemical and biological river quality data. NRA R&D Note 198.

Mardia, K.V., Kent, J.T. and Bibby, J.M. (1979) *Multivariate analysis*. Academic Press, London.