WRc

National Rivers Authority Anglian Region

Regional Operational Investigation 579

Quantification of the relationship between effluent quality and biological quality

Interim Report for Stage 1

WRc Ref: CO4093
JANUARY 1996

Regional Operational Investigation 579

QUANTIFICATION OF THE RELATIONSHIP BETWEEN EFFLUENT QUALITY AND BIOLOGICAL QUALITY: INTERIM REPORT FOR STAGE 1.

A Gunby, I Milne and M Wheeler

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

National Rivers Authority Anglian Region

National Rivers Authority Anglian Region
Kingfisher House Golday Way
Orton Golday, Peterborough PE2 5ZR

Dissemination status
Internal: Limited Release
External: Restricted

Statement of use
This report is the first output from Regional Operational Investigation 579. It covers Stage 1 of the project, over the period February 1995 to January 1996. It is to be used for information on progress to date.

Research contractor
This document was produced under Contract by:

WRc plc
Henley Rd Medmenham
Marlow
SL7 2HD

Tel: 01491 571531
Fax: 01491 579094

WRc Report Nº CO4093/09106

January 1996

Contract Manager: I Milne

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

Dr S Chadd

CONTENTS Page

Page

LIST OF TABLES

LIST OF FIGURES

Page

Page

1. INTRODUCTION

This Interim Report is the first formal output 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 NRA 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 are:

Stage 1:

1. To collate existing details and data from sources within the NRA 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.

The project commenced in February 1995, with a start-up meeting at which the aims and objectives of the project were discussed and possible approaches to data analysis were explored. This report represents the completion of Stage 1.

2. DATA RECEIVED

Data requirements were identified and agreed at the start-up meeting, and the data was supplied to WRc in electronic format. The supplied data derived from a number of NRA databases and comprised the following:

- Details of 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)

3. 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 NRA-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 NRA 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 NRA 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.

Table 3.1 shows that of a total of 328 STWs in the Relate database 310 had both upstream and downstream biological monitoring sites.

Table 3.1 Number of STWs in 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 3.2 shows the number of STWs with associated biological and other data.

Table 3.2 Number of STWs 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

Clearly the number of STWs with all necessary data available for both upstream and downstream, as appropriate, is limited. Table 3.3 shows the progressive reduction of the size of the available dataset as different data components are introduced. The final number of STWs with all available data is 29, although it should be noted that for some of these sites some of the data is limited.

Table 3.3 Number of STWs 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

4. PRELIMINARY DATA ANALYSIS

4.1 Introduction

Early in the project a preliminary analysis was undertaken with the data available at the time. This data set did not include river flow information, data on the chemical quality of the receiving waters or RIVPACS predictions for the receiving waters. Without these determinands there was too much unexplained variation in the data to be able draw any useful conclusions, although there did appear to be a relationship between the biological scores and the effluent chemistry determinands.

The available biological determinands were:

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

The effluent determinands available were:

- total population equivalent of the STW,
- dry weather flow of the STW,
- effluent BOD concentrations (mg l^{-1}),
- effluent ammonia (NH_4) concentrations (mg l^{-1}), and
- effluent suspended solids (SS) concentrations (mg l^{-1}).

4.2 Summary statistics

Table 4.1 below shows simple summary statistics for the transformed biological variables. These statistics show that the differences between downstream and upstream biological sites were, on average, close to zero (although all three measures have means which were just less than zero).

There appeared to be several STWs which showed a large increase in biological quality indicators downstream of the discharge. However, it was not possible at that point in the study to determine how many of these were due to anomalies in the database or to natural differences between biological sampling sites.

Table 4.1 Summary statistics for the transformed biological variables

Statistic	Difference between downstream and upstream variable		
	BMWP	ASPT	LQI
Mean	-4.86	-0.06	-0.14
s.d.	42.2	0.94	1.32
CoV	-8.69	-16.7	-9.43
Maximum	137	5.15	3.50
Median	-4.50	-0.04	0.00
Minimum	-167	-5.16	-4.50

The mean and standard deviation (s.d.) of the transformed biological variables within each season are given in Table 4.2. The seasons were defined to be

Spring - February to May inclusive,

Summer - June to August inclusive, and

Winter - September to January inclusive.

Table 4.2 Summary statistics for the transformed biological variables split by season

Season	Statistic	Difference between downstream and upstream variable		
		BMWP	ASPT	LQI
Spring	Mean	-2.74	0.02	0.05
	s.d.	41.10	0.80	1.35
Summer	Mean	-8.30	-0.10	-0.30
	s.d.	34.97	0.71	1.15
Winter	Mean	-3.17	-0.02	-0.14
	s.d.	46.55	0.89	1.49

It would seem reasonable to assume that the STWs would generally have a greater impact on river quality during the Summer months, since this is when dilution is likely to be least, and higher water temperatures may exacerbate the impact associated with oxygen demand. This is perhaps reflected by the slight widening in mean difference during Summer, but the size of the difference is small when compared with the amount of variability; analysis of variance of any of these biological variables on the seasonal factor showed no significant effects.

Table 4.3 gives the summary statistics for the chemical variables for reference, and Table 4.4 gives the means and standard deviations within seasons. There was no discernible seasonal variation in the chemical determinands. Note, however, that, as these figures are means for all STWs, seasonality at individual STWs may have been obscured.

Table 4.3 Summary statistics for the log-transformed chemical determinands

Statistic	Log of determinand				
	TPE	DWF	BOD	NH_4	SS
Mean	8.12	6.73	2.12	0.51	2.61
s.d.	1.73	1.71	0.61	1.10	0.59
CoV	0.21	0.25	0.29	2.14	0.23
Minimum	3.40	3.47	0.02	-1.90	1.18
Median	8.12	6.49	2.15	0.59	2.62
Maximum	12.7	10.8	4.74	3.86	5.06

Table 4.4 Summary statistics for transformed chemical determinands split by season

Season	Statistic	Log of determinand				
		TPE	DWF	BOD	NH_4	SS
Spring	Mean	8.15	6.72	2.11	0.51	2.60
	s.d.	1.70	1.71	0.58	1.10	0.59
Summer	Mean	8.12	6.72	2.16	0.56	2.63
	s.d.	1.71	1.68	0.64	1.12	0.62
Winter	Mean	8.07	6.69	2.12	0.52	2.61
	s.d.	1.8	1.7	0.62	1.07	0.58

4.3 Canonical correlation analysis

A canonical correlation (see Section 6.2 for description of this technique) was undertaken using the biological data, transformed to the difference between downstream and upstream scores, as one set of variables and the effluent determinands, log-transformed, as the other set.

Table 4.5, below, shows the correlations of the canonical variates and their percentages of the overall correlation. The correlations are quite small (even the 1st variate only has a correlation of 0.33) and the first two variates only explain 80.9% of the total correlation.

Table 4.5 Correlations of the canonical variates

Canonical variate	Correlation	% correlation	Cumulative % correlation
1	0.33	52.3	52.3
2	0.18	28.6	80.9
3	0.12	19.1	100

The amount of variation in the biological variables (differences between downstream and upstream) explained by the physical variables is obviously very little (about 15%) and so any relationship present in this data set will have little power in predicting changes in biology. In addition to this, not all of this 15% of the variation explained is due to the effluent quality data.

5. ASSESSMENT OF IMPORTANT FACTORS

5.1 Introduction

Following the auditing and revision of the database (See Section 3), more detailed statistical analysis was undertaken of the data from the 29 STWs for which the full set of determinands was available. This analysis incorporated the determinands that had not been available for the earlier analysis documented in Section 4. The determinands available in the final database 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.

5.2 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. 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).

All other things being equal, the former of these two measure may be favoured, as it is already used by the NRA as an Ecological Quality Index (EQI).

To compare the biological quality of the downstream sites with those of the upstream, for case a) above, the ratio of downstream to upstream EQI was used (e.g. BMWP EQI downstream / BMWP EQI upstream). However, in case b) above, the ratio of the differences between observed and predicted biology is not defined when the upstream biological score equals the prediction (giving a zero on the denominator). Therefore the difference between the downstream difference and the upstream difference was used instead, for example, (BMWP - RIVPACS downstream) - (BMWP - RIVPACS upstream). For LQI the difference between downstream and upstream sites was used in all cases.

5.3 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 $x \times DWF \times 1000$ 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 $y \times MRF \times 1000$ 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).

5.4 Examination of possible relationships

5.4.1 EQI and theoretical increases in river chemistry

Figures 5.1 - 5.4 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. The points in all of these figures are scattered around a ratio of 1. This seems to indicate that there is no relationship between the theoretical increase in receiving water concentration and changes in biology downstream. One site has been singled out as being a possible outlier: STW S01LETCHTHDP (marked as hollow squares on the figures) has a dry weather flow rate of about 7 times that of the river it discharges into.

One possible reason why there may not be an apparent relationship is because of the limited range of theoretical increases in receiving water concentrations. In other word, the STWs examined were not generally having a significant impact on receiving water chemistry and therefore not on receiving water biology. For BOD, most of the increases are below 0.5 mg l^{-1} and for ammonia they are mostly less than 0.25 mg l^{-1}. These small increases, coupled with the inherent variability in the EQI measure (even where there is no theoretical worsening of river quality the ratio of BMWP EQIs ranges between 0.3 and 2), may mean that there is in fact a relationship but that data are needed from STWs where the theoretical increase in receiving water chemistry is substantially greater in order to detect it.

Figures 5.5-5.8 show the equivalent plots but using the alternative measure of change in biology, i.e. the difference between downstream and upstream differences in observed and predicted biology scores. These plots show generally the same pattern as for the EQIs, and do not appear to improve the relationship.

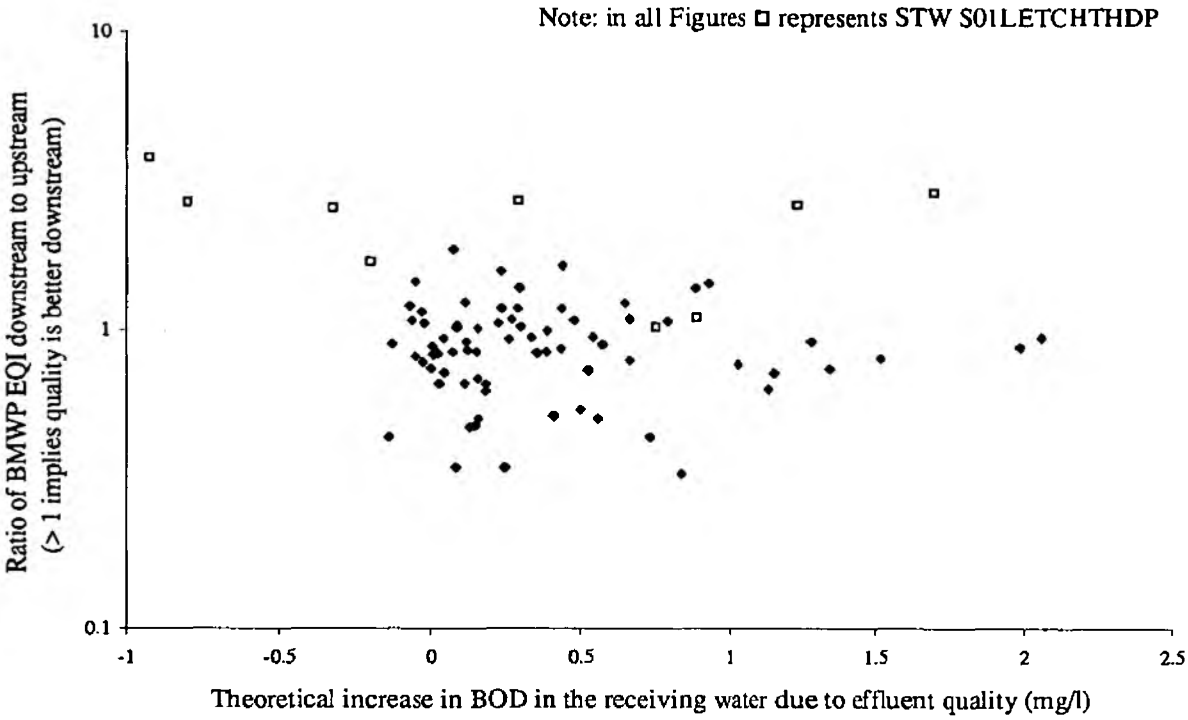

Figure 5.1 Relationship between BMWP EQI ratio and theoretical BOD increase

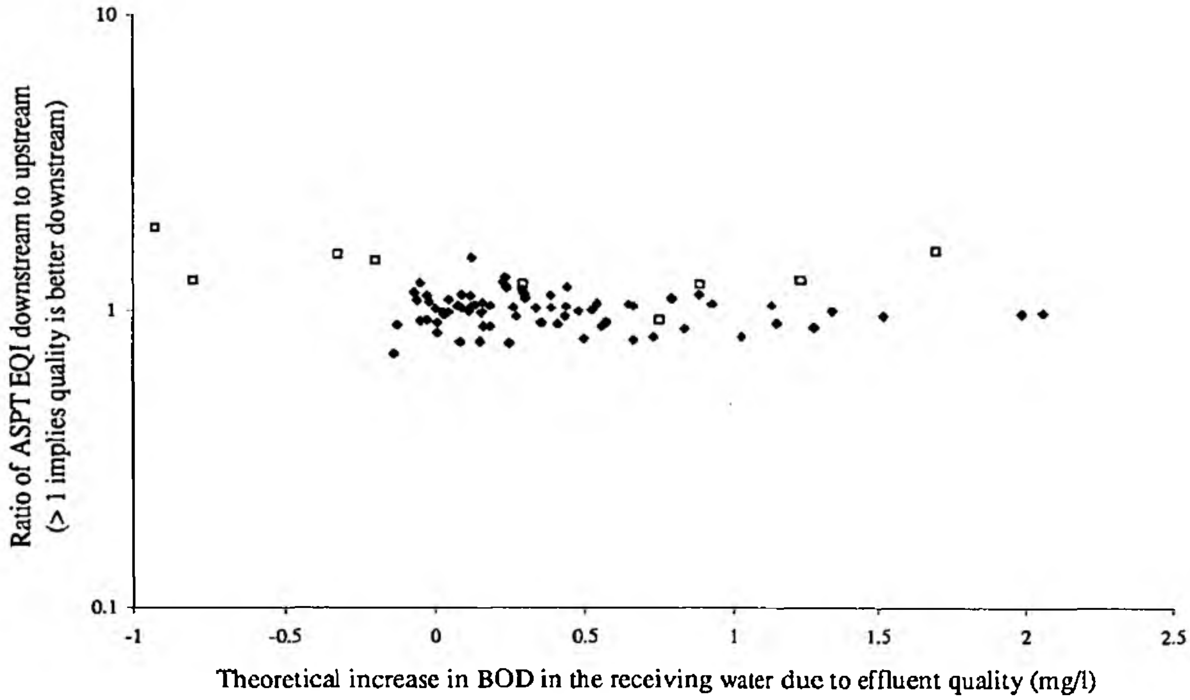

Figure 5.2 Relationship between ASPT EQI ratio and theoretical BOD increase

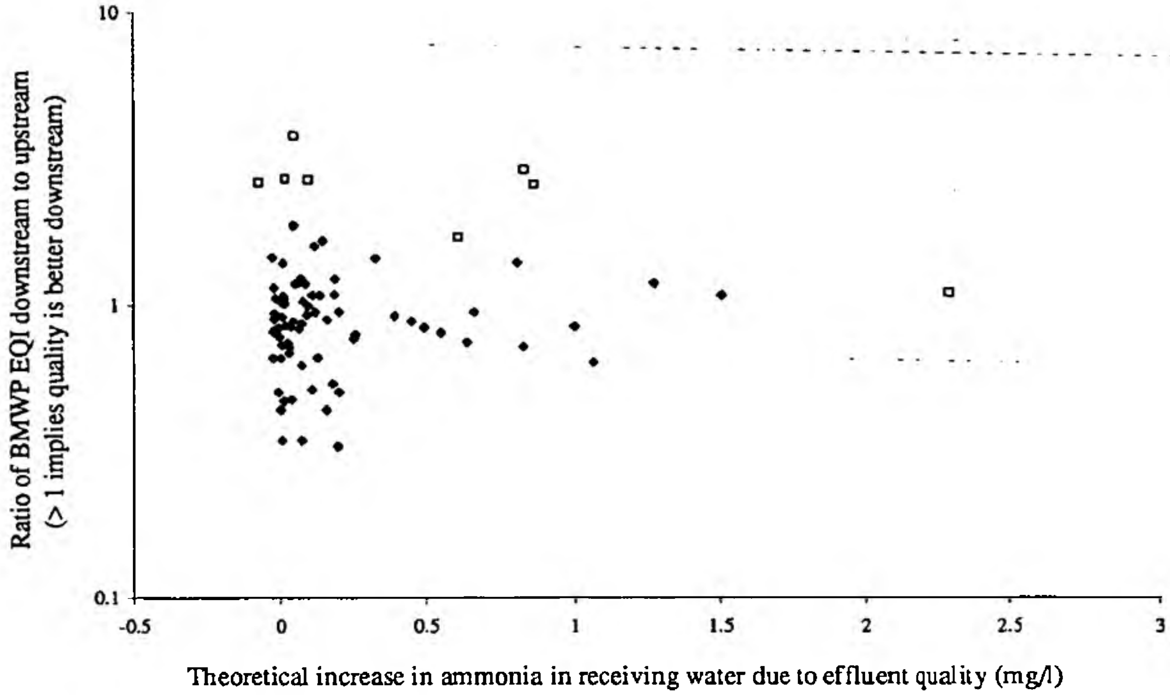

Figure 5.3 Relationship between BMWP EQI ratio and theoretical ammonia increase

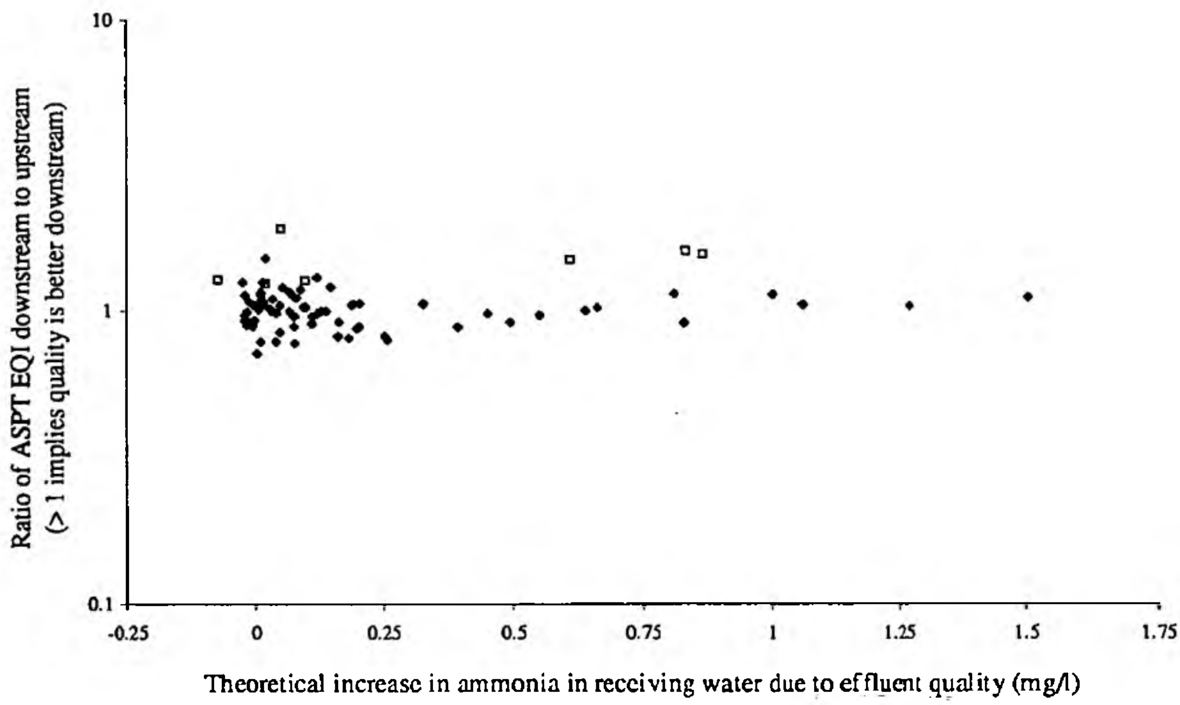

Figure 5.4 Relationship between ASPT EQI ratio and theoretical ammonia increase

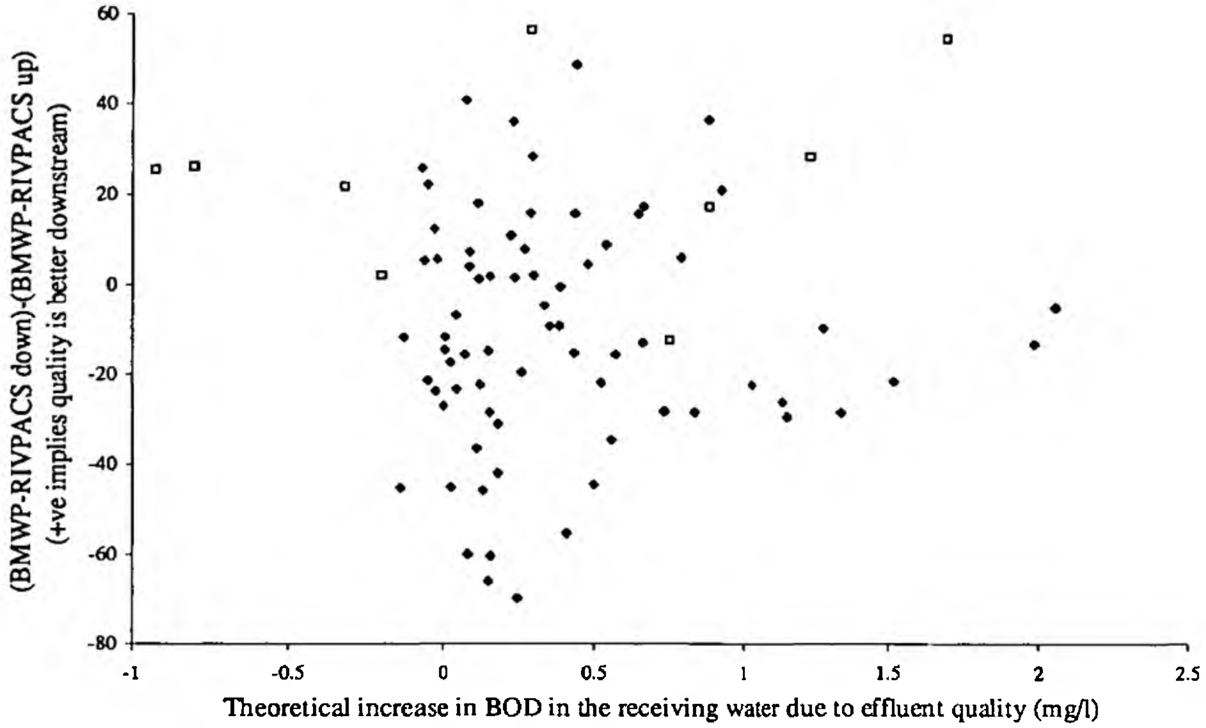

Figure 5.5 Relationship between BMWP-RIVPACS difference and theoretical BOD increase

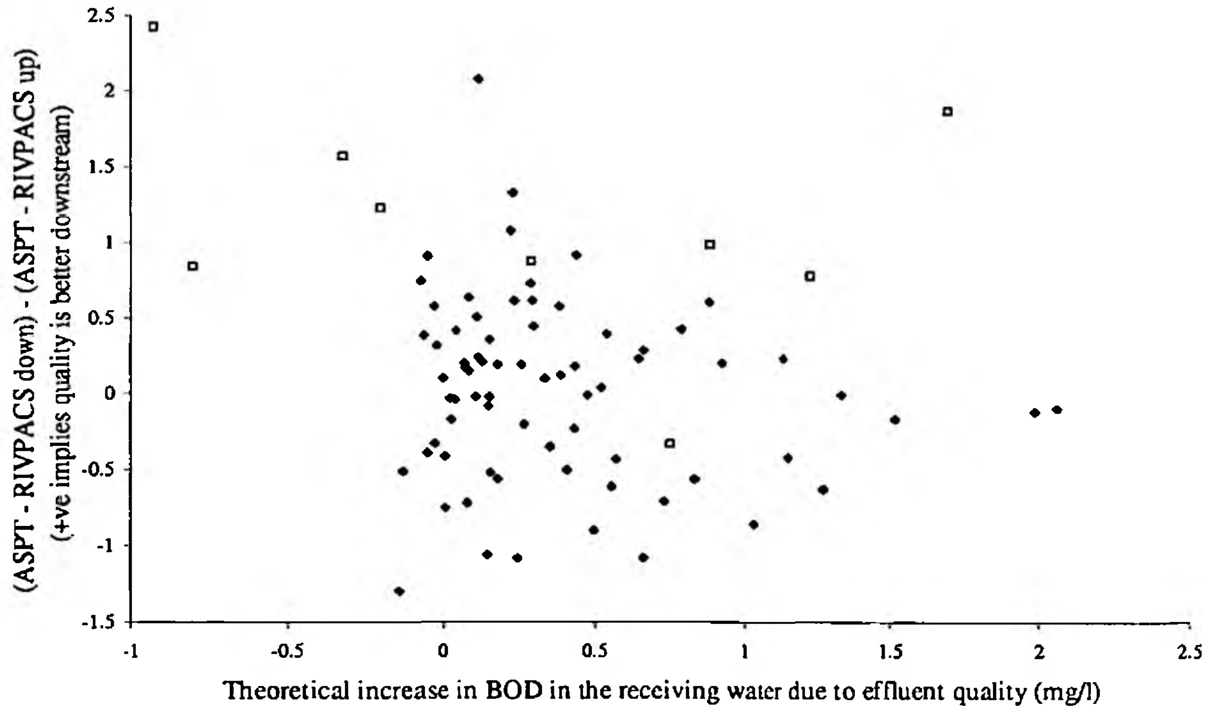

Figure 5.6 Relationship between ASPT-RIVPACS difference and theoretical BOD increase

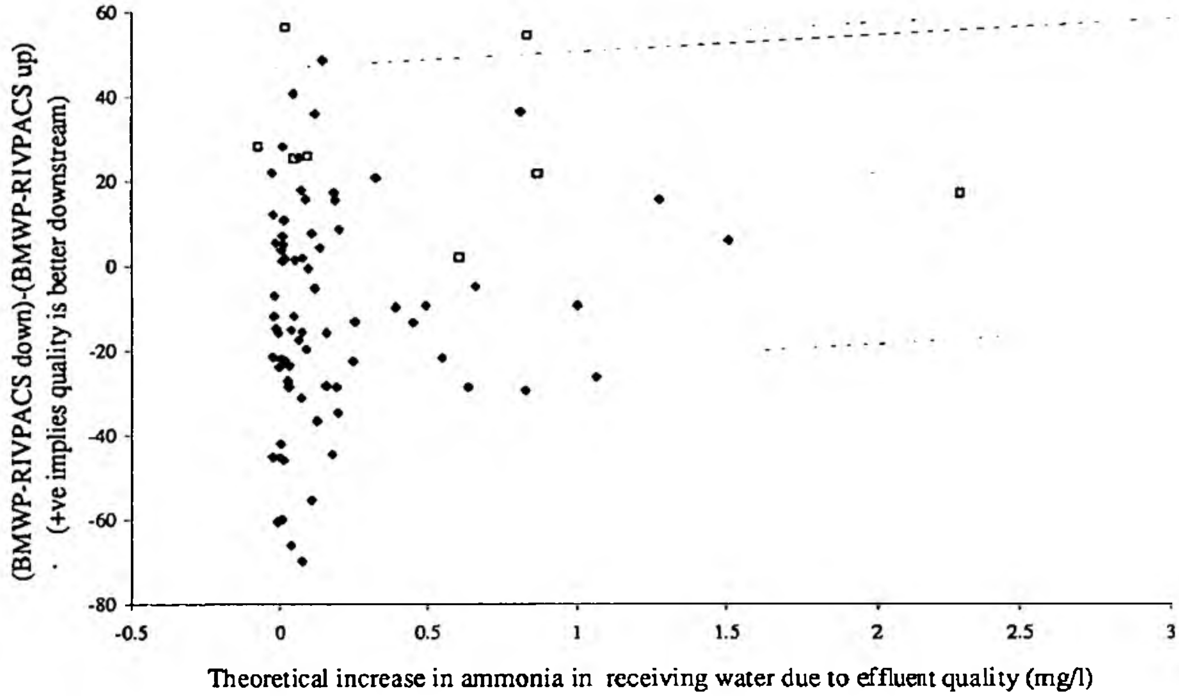

Figure 5.7 Relationship between BMWP-RIVPACS difference and theoretical ammonia increase

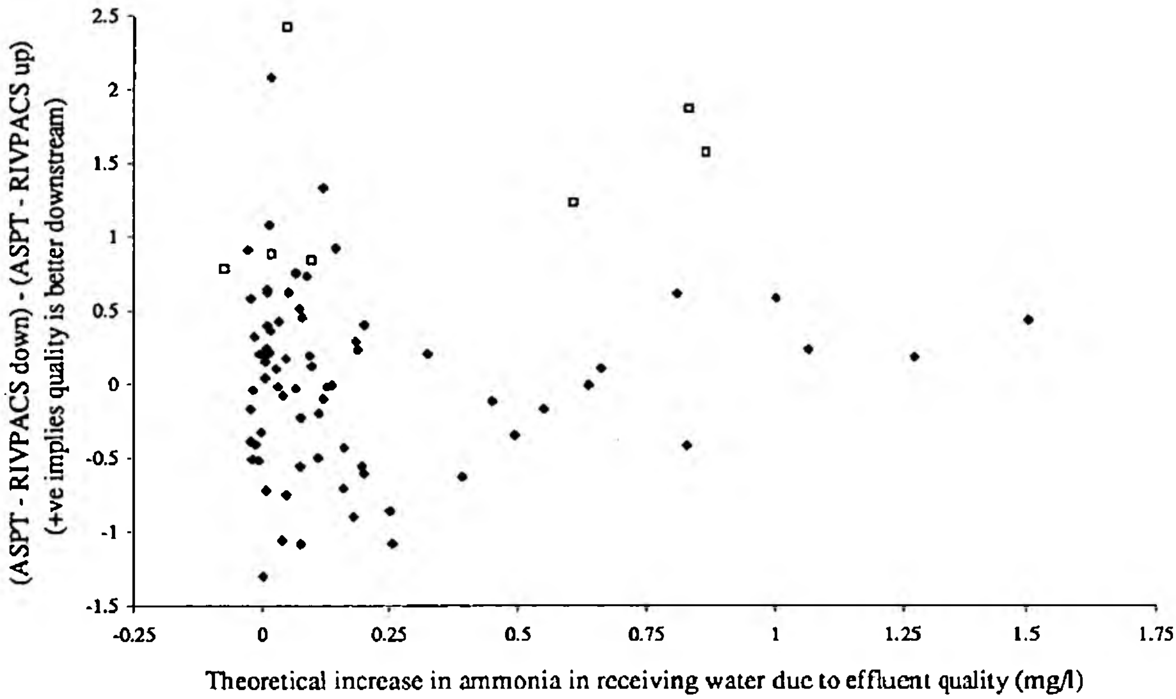

Figure 5.8 Relationship between ASPT-RIVPACS difference and theoretical ammonia increase

5.4.2 EQIs and changes in GQA means

In Figures 5.9-5.14 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 relationship between chemical quality and biology to be revealed in these figures. However, as with the theoretical increases in receiving water concentrations in Figures 5.1 - 5.4, the changes in GQA means are mostly limited to a range about 0.5 mg l^{-1} wide (although for dissolved oxygen the range is about 20 percentage points) with a few clusters of points lying relatively far away. These extreme groups have an undue influence on the direction of the regression lines fitted to these data. Therefore one must be cautious about drawing any conclusions from these plots about the nature of the true relationship between changes in biology and river chemistry.

These plots are repeated for the difference-based changes in biology in Figures 5.15-5.20. As before, the use of this alternative measure of biological change does not show any apparent difference from the EQI plots.

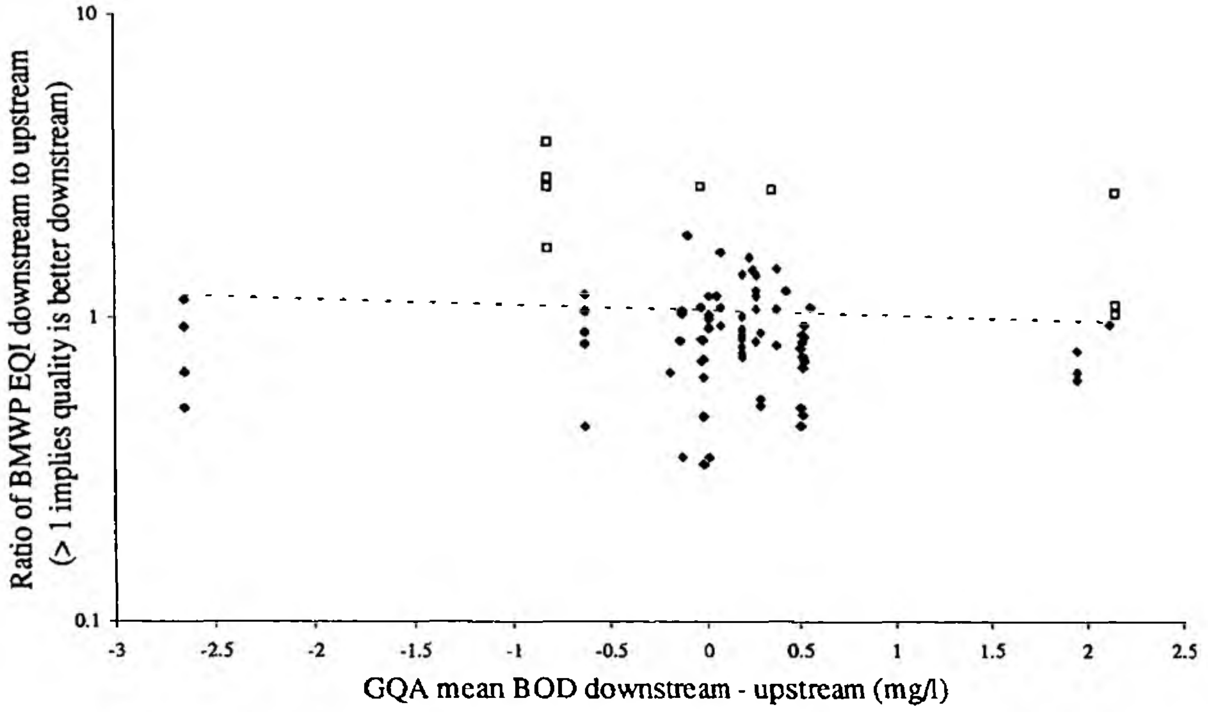

Figure 5.9 Relationship between BMWP EQI ratio and change in GQA BOD

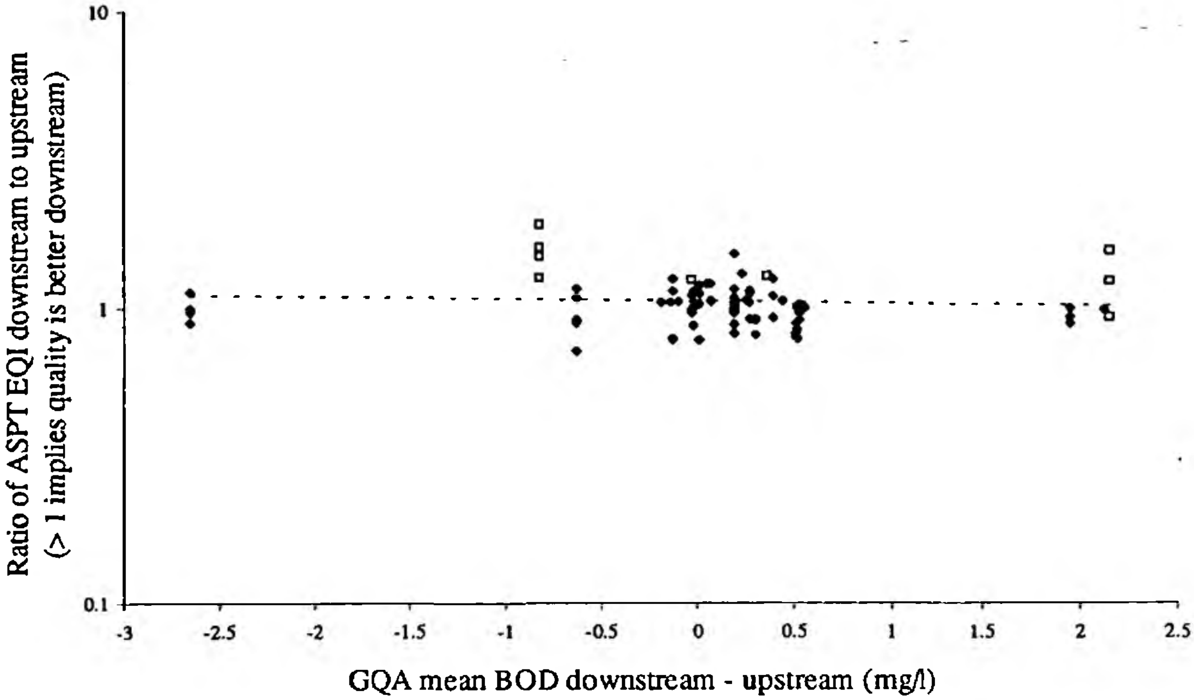

Figure 5.10 Relationship between ASPT EQI ratio and change in GQA BOD

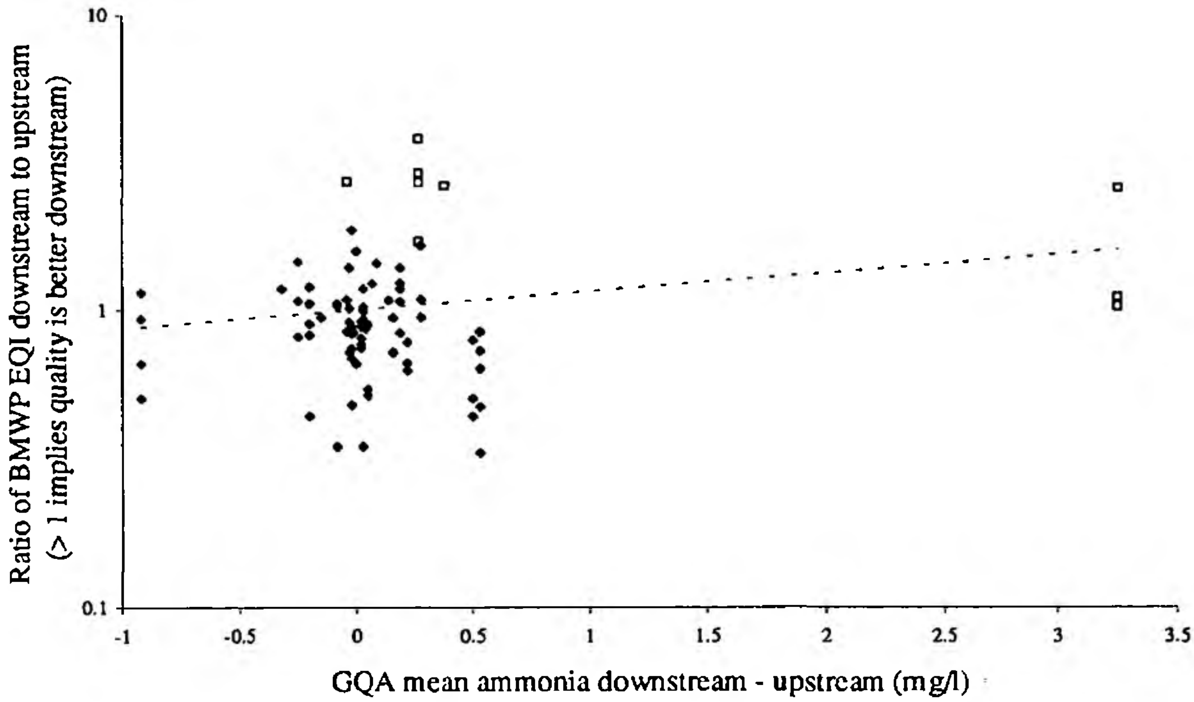

Figure 5.11 Relationship between BMWP EQI ratio and change in GQA ammonia

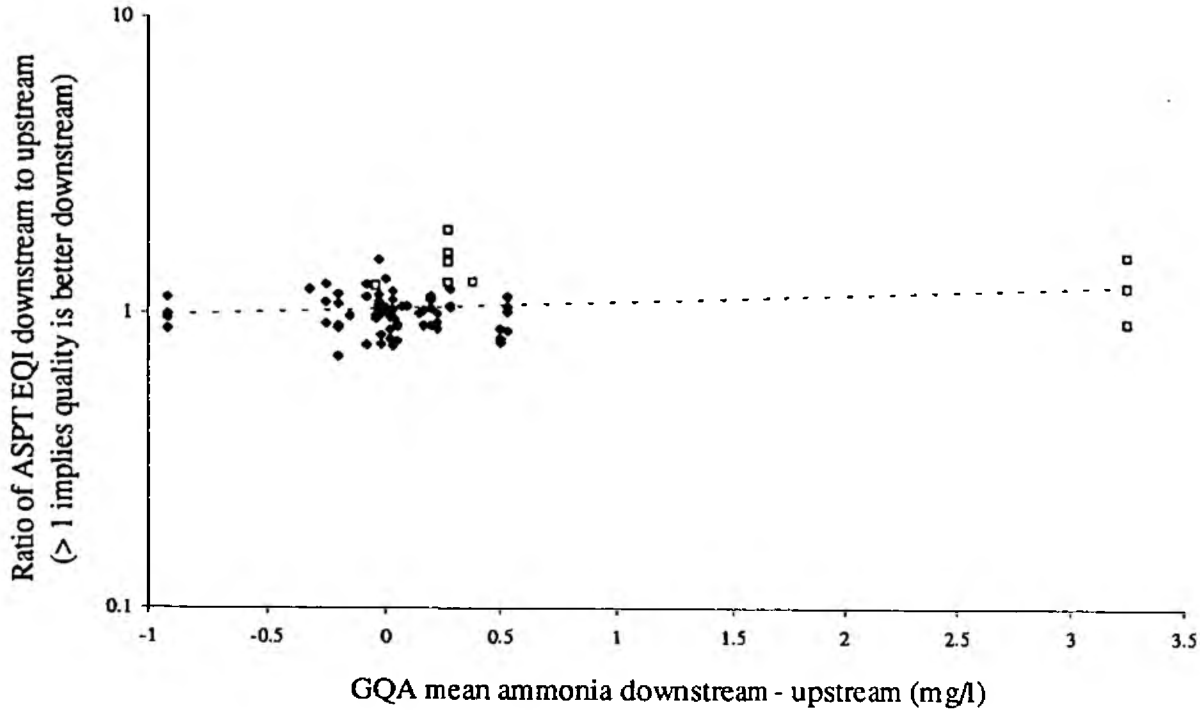

Figure 5.12 Relationship between ASPT EQI ratio and change in GQA ammonia

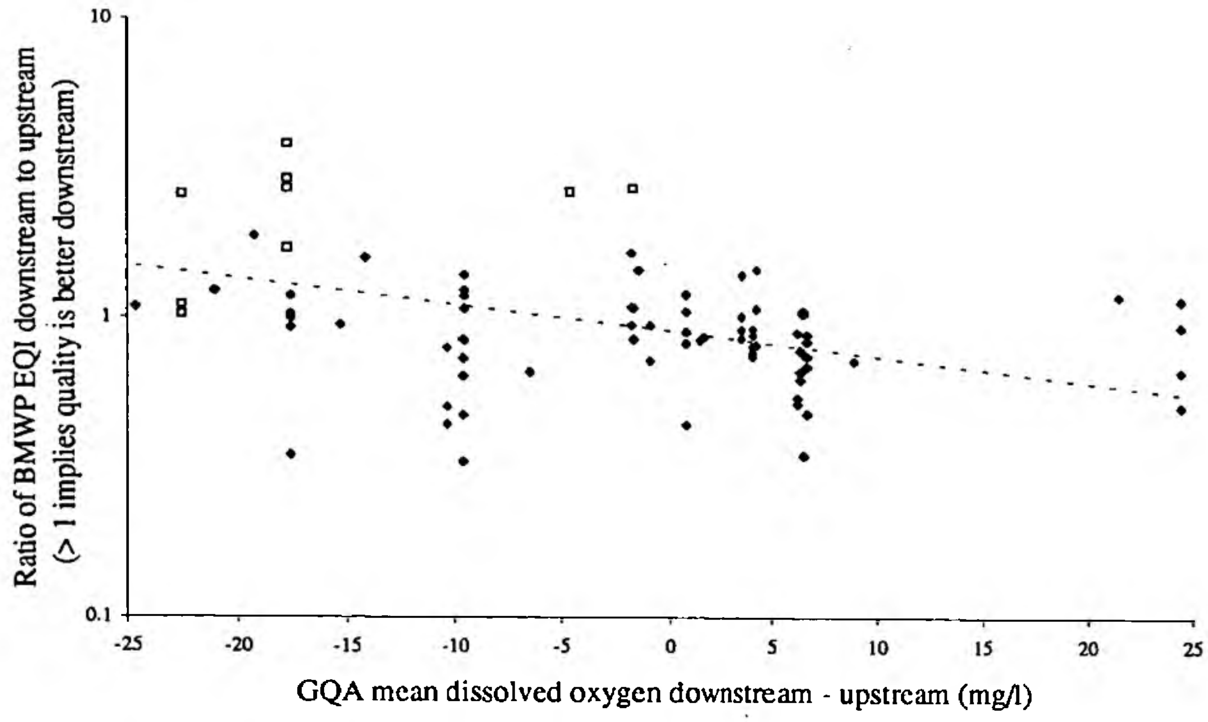

Figure 5.13 Relationship between BMWP EQI ratio and change in GQA dissolved oxygen

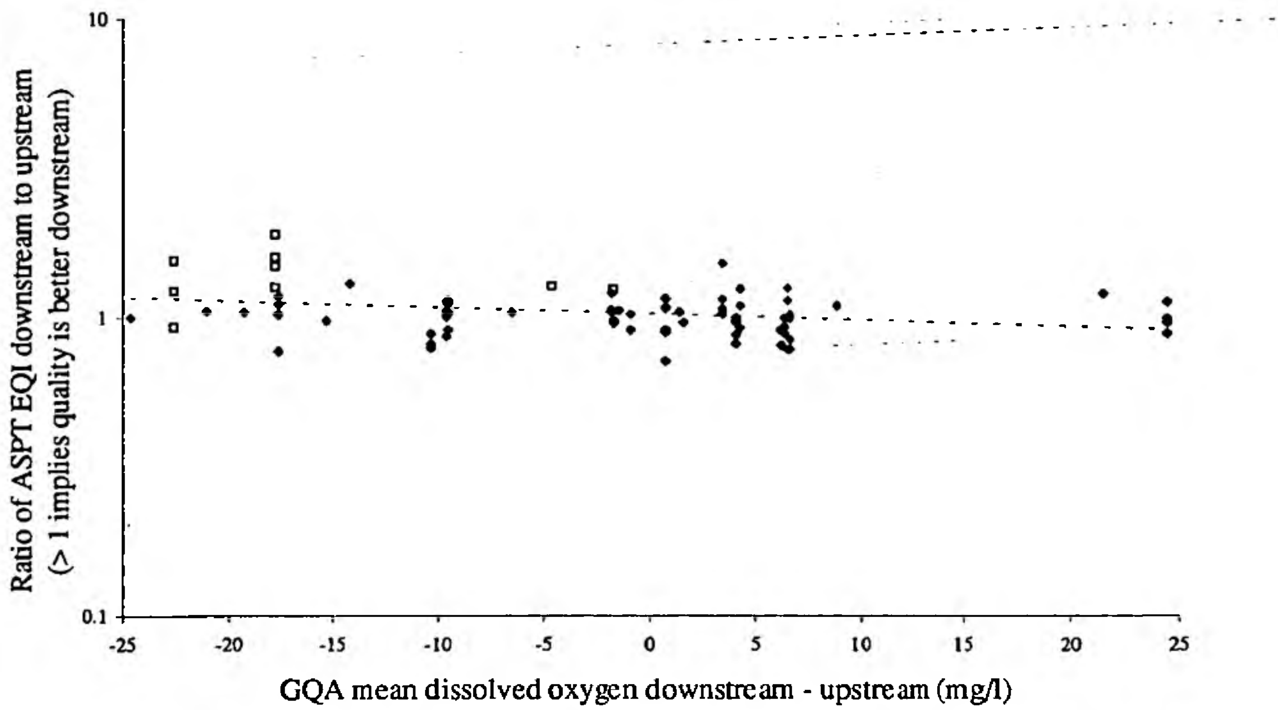

Figure 5.14 Relationship between ASPT EQI ratio and change in GQA dissolved oxygen

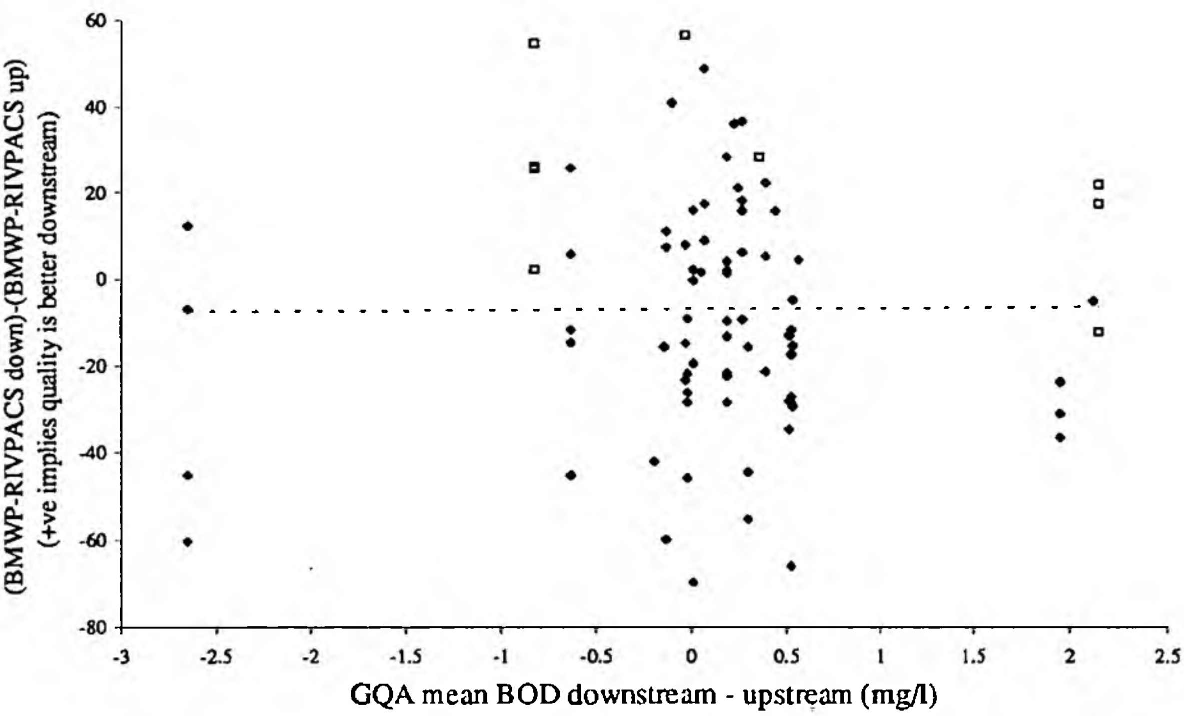

Figure 5.15 Relationship between BMWP-RIVPACS difference and change in GQA BOD

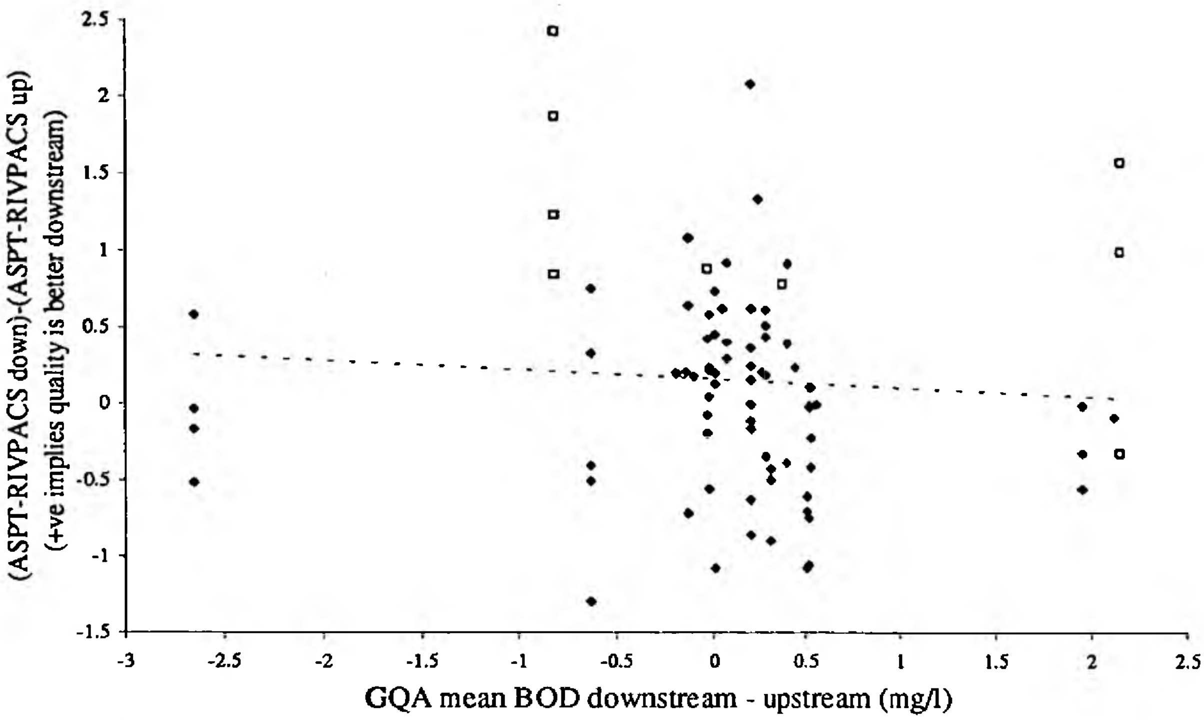

Figure 5.16 Relationship between ASPT-RIVPACS difference and change in GQA BOD

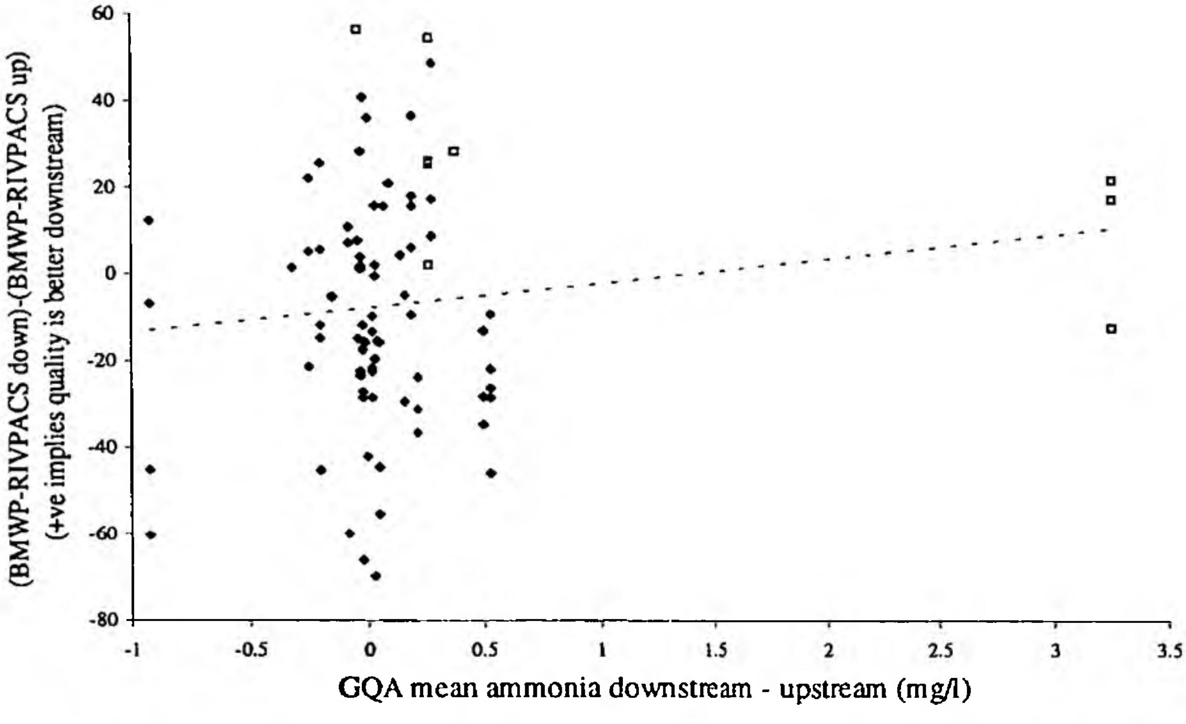

Figure 5.17 Relationship between BMWP-RIVPACS difference and change in GQA ammonia

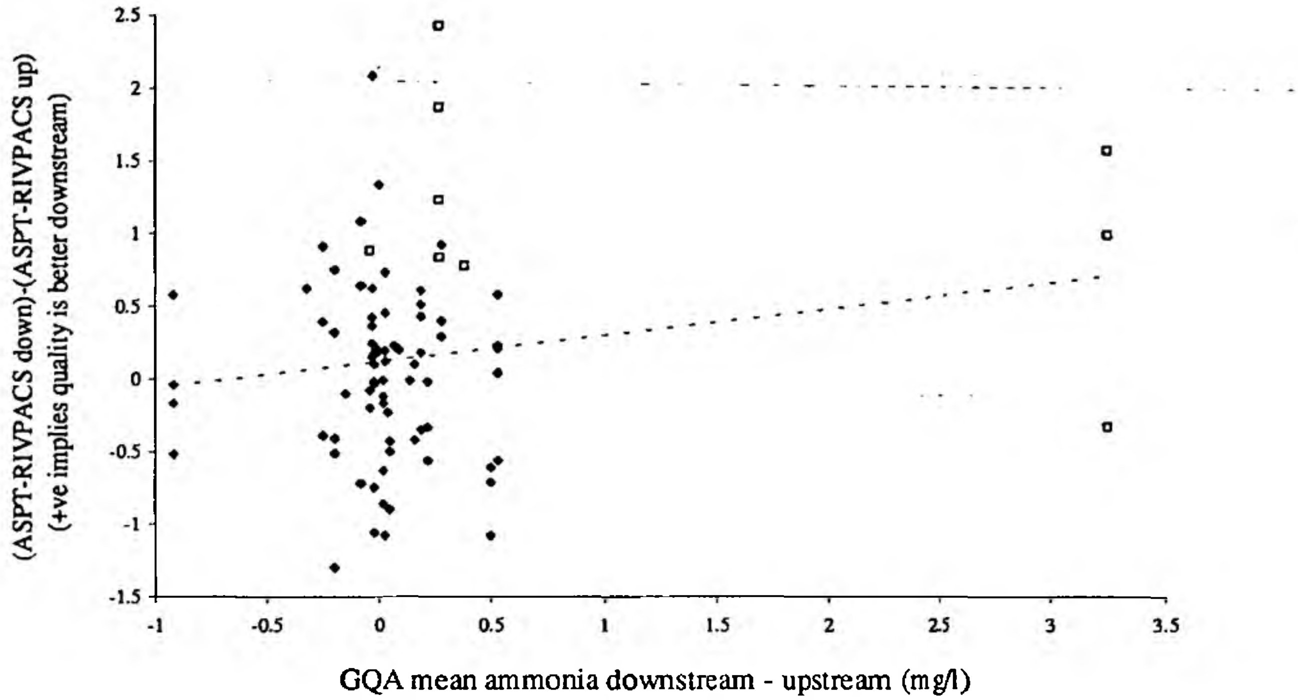

Figure 5.18 Relationship between ASPT-RIVPACS difference and change in GQA ammonia

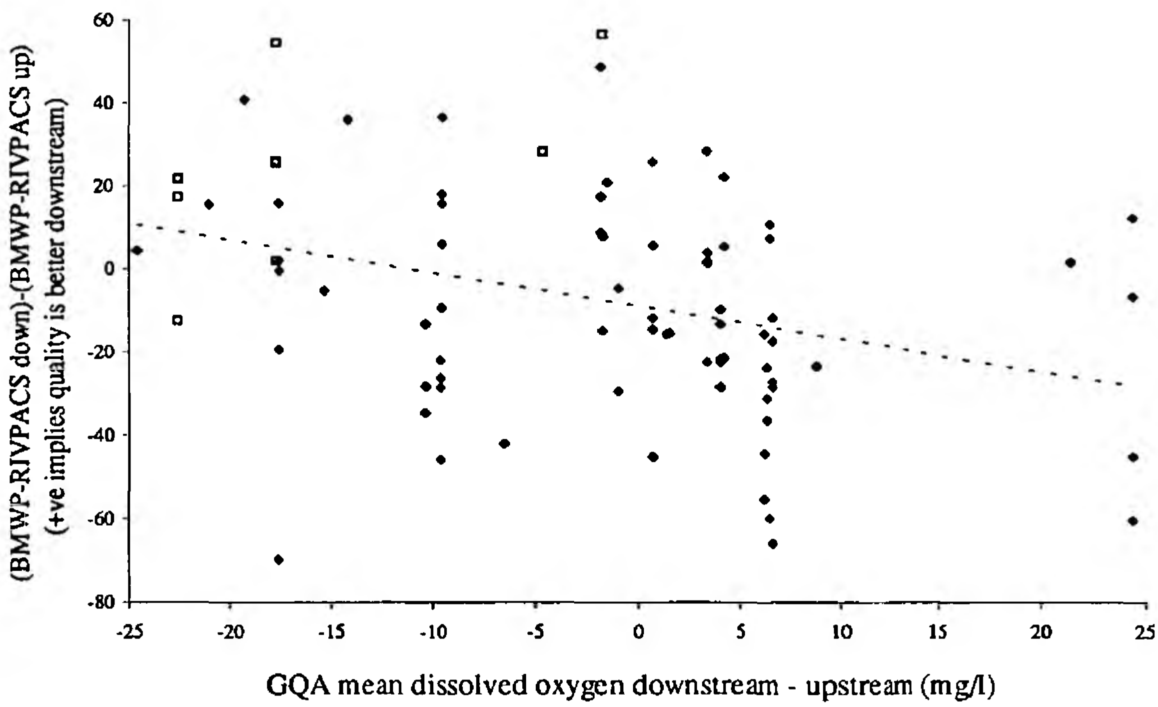

Figure 5.19 Relationship between BMWP-RIVPACS difference and change in GQA dissolved oxygen

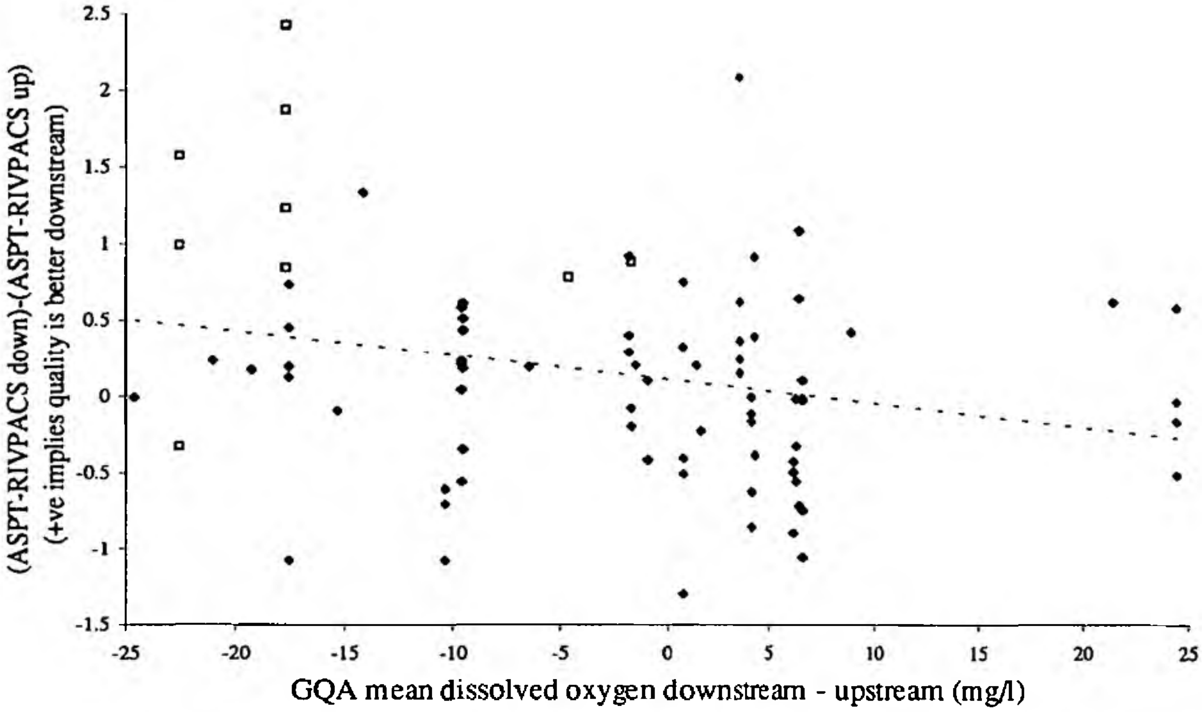

Figure 5.20 Relationship between ASPT-RIVPACS difference and change in GQA dissolved oxygen

5.4.3 Influence of RIVPACS predictions

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, 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. Plots of some of these relationships are shown in Figures 5.21-5.24.

Inspection of Figures 5.21-5.24 reveals that there is no apparent improvement in the relationship between biology and chemistry when RIVPACS predictions are excluded.

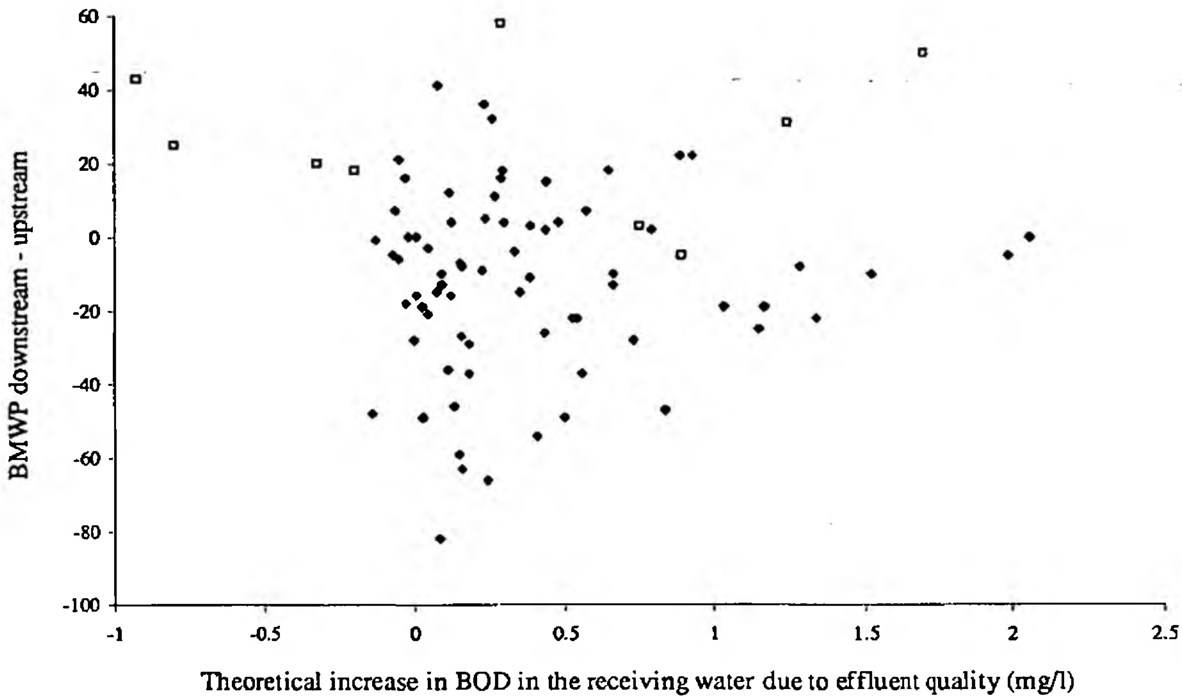

Figure 5.21 Relationship between BMWP difference and theoretical increase in BOD

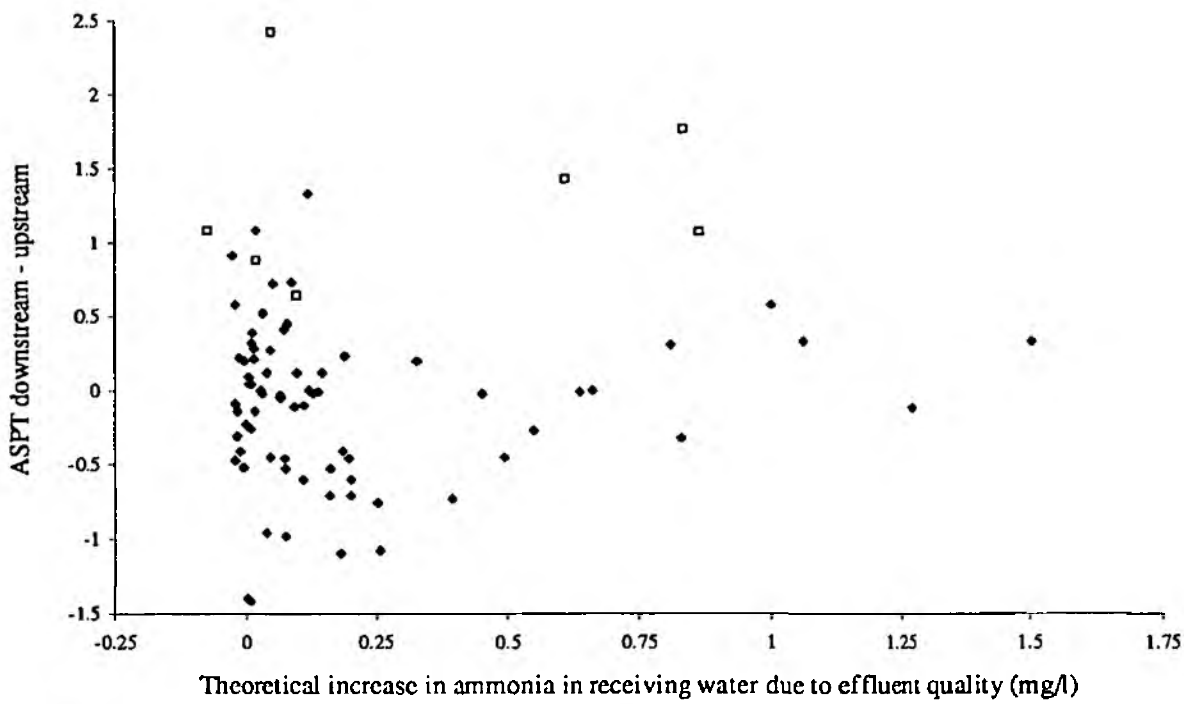

Figure 5.22 Relationship between ASPT difference and theoretical increase in ammonia

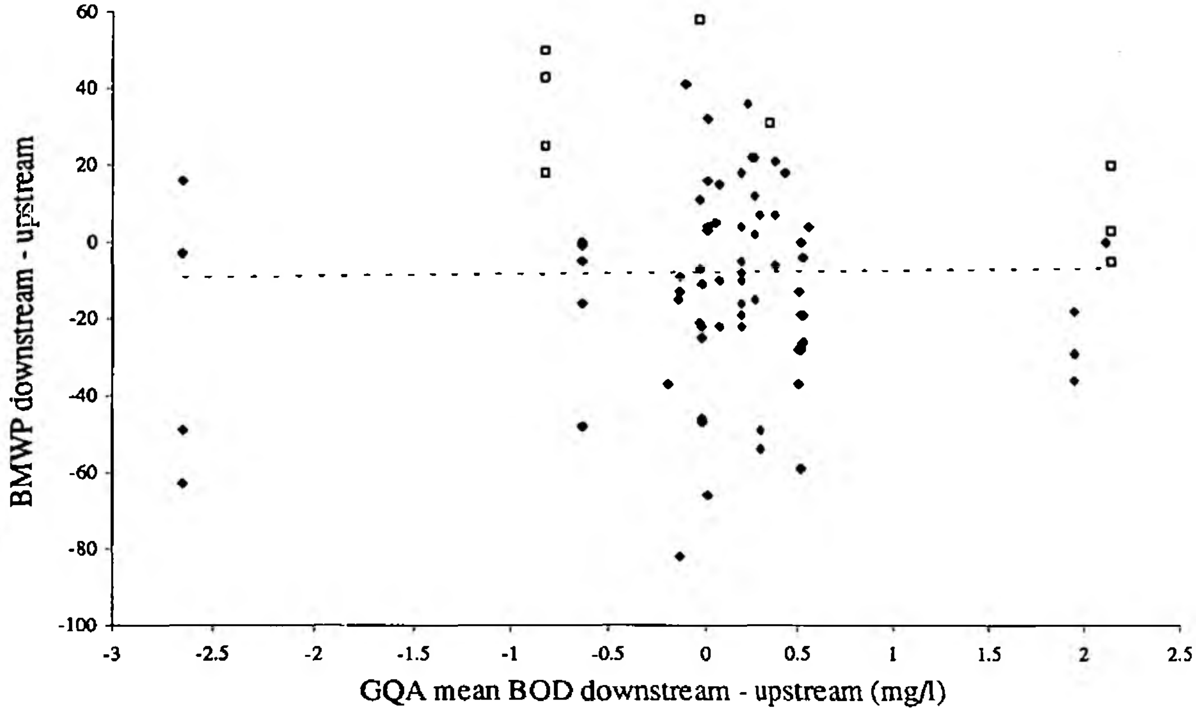

Figure 5.23 Relationship between BMWP difference and change in GQA ammonia

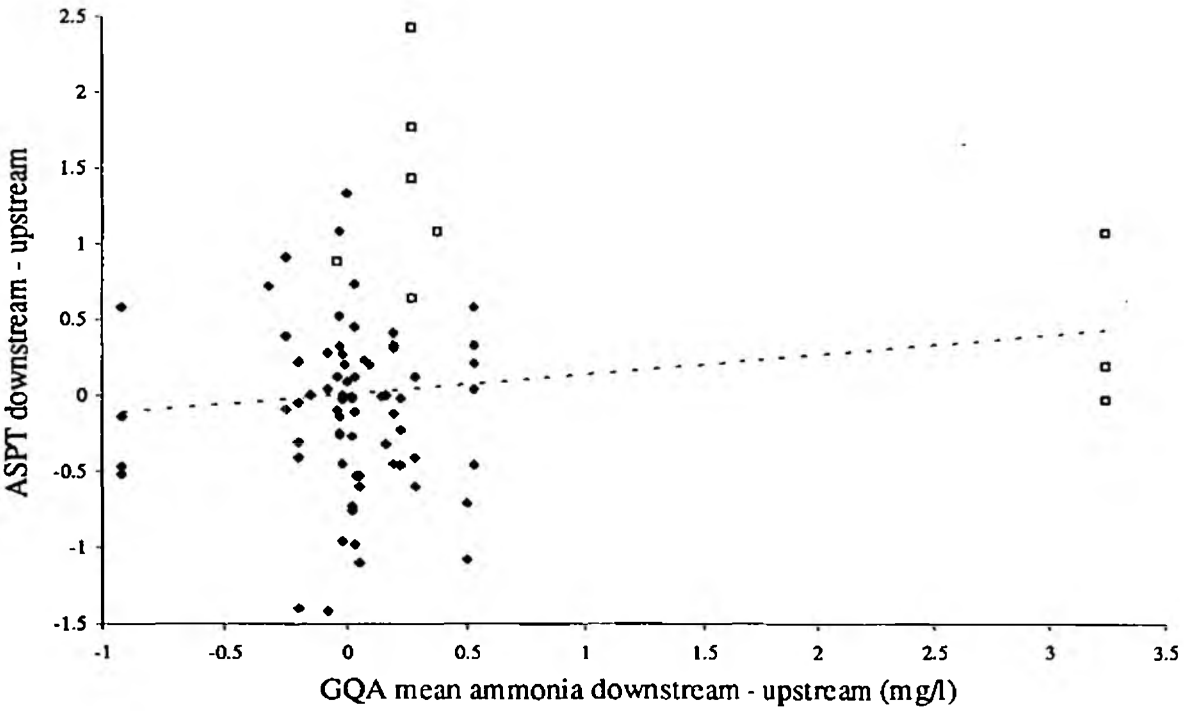

Figure 5.24 Relationship between ASPT difference and change in GQA ammonia

5.4.4 GQA means and theoretical increase in river chemistry

Figures 5.25 and 5.26 show plots of the changes in GQA means for BOD and ammonia against the theoretical changes in river chemistry due to the discharges. As with the previous plots there is not enough data spread along a large enough range of values to see any relationship. One would hope that there would be a good correlation between the two variables. The fact that there is not may indicate that there is too much variability in the data and so more samples are required.

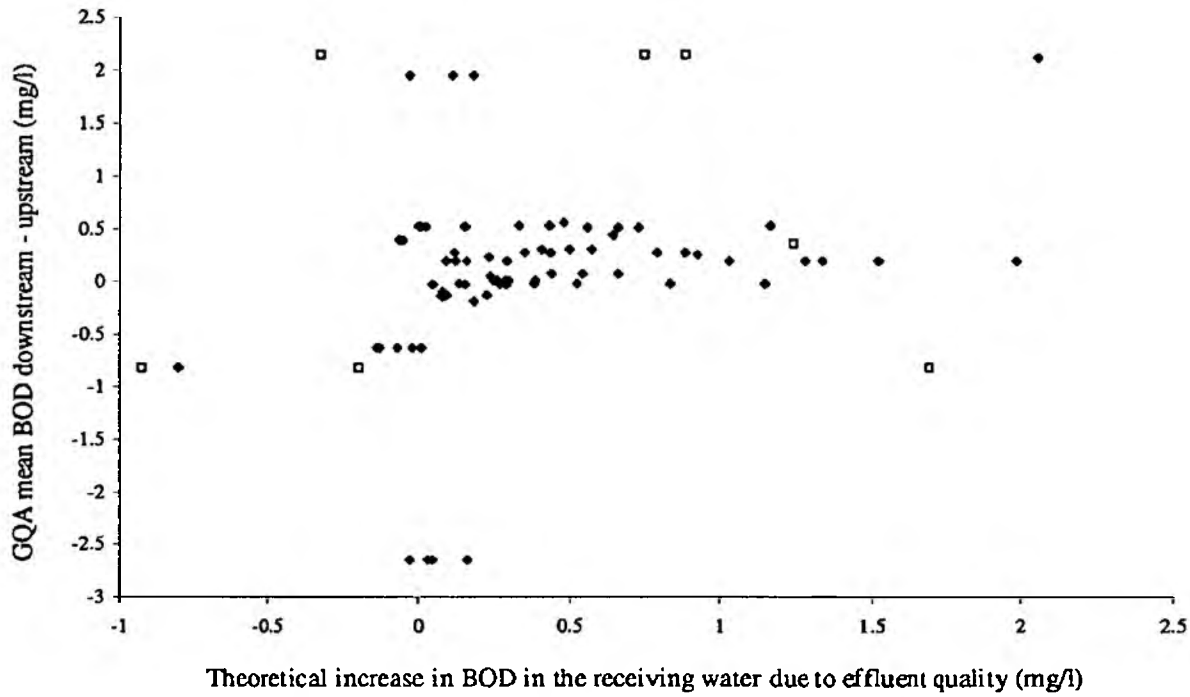

Figure 5.25 Relationship between GQA change and theoretical increase for BOD

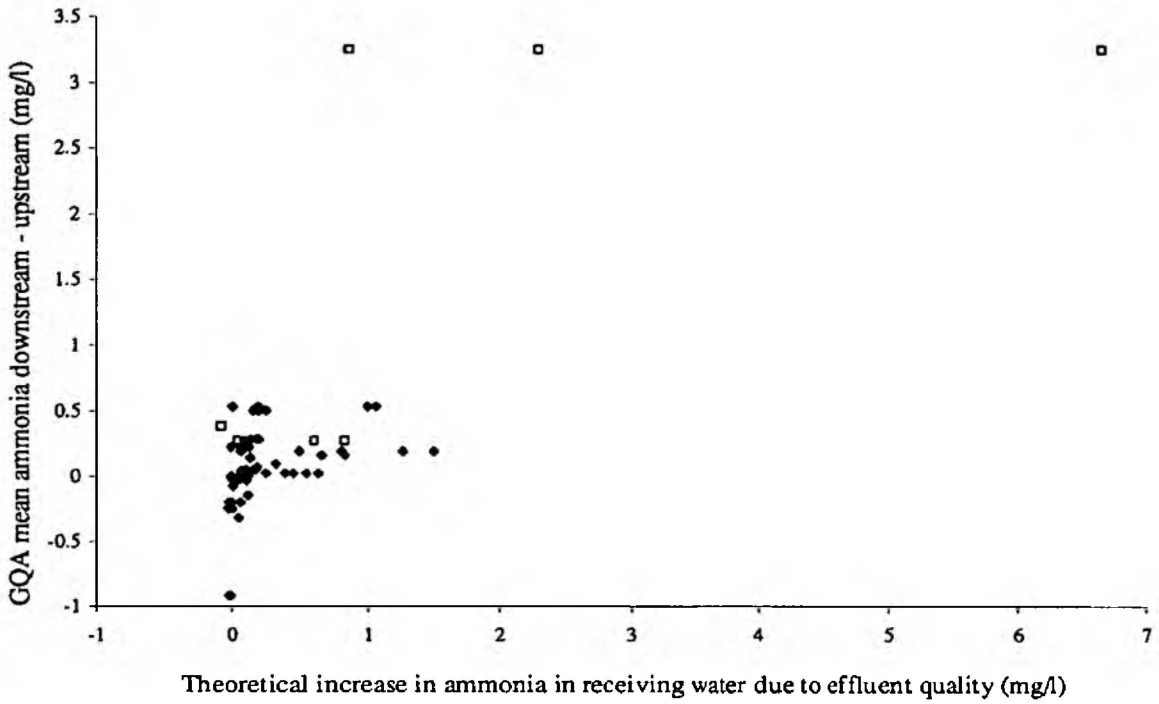

Figure 5.26 Relationship between GQA change and theoretical increase for ammonia

5.4.5 Changes versus time

In Figure 5.27 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 couple of STWs most of these theoretical increases remain fairly constant through time. This is also repeated in Figure 5.28 where the differences in GQA mean BOD are plotted. The plot of the ratio of ASPT EQIs against time in Figure 5.29 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.

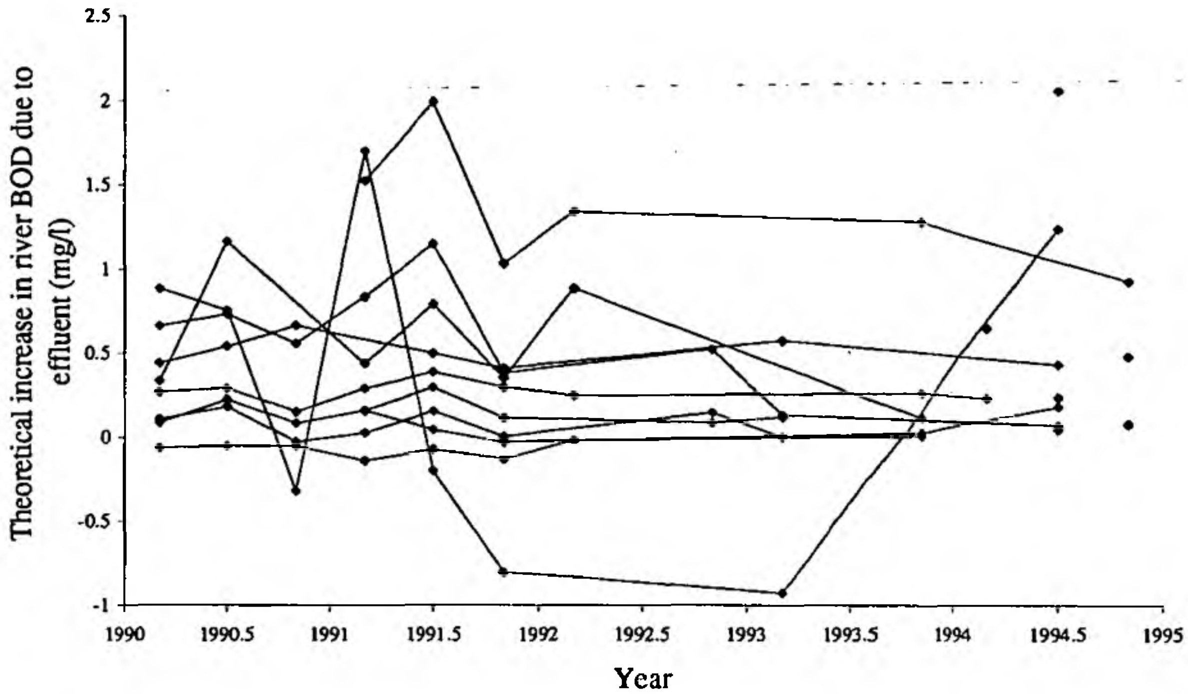

Figure 5.27 Theoretical increase in BOD plotted against time

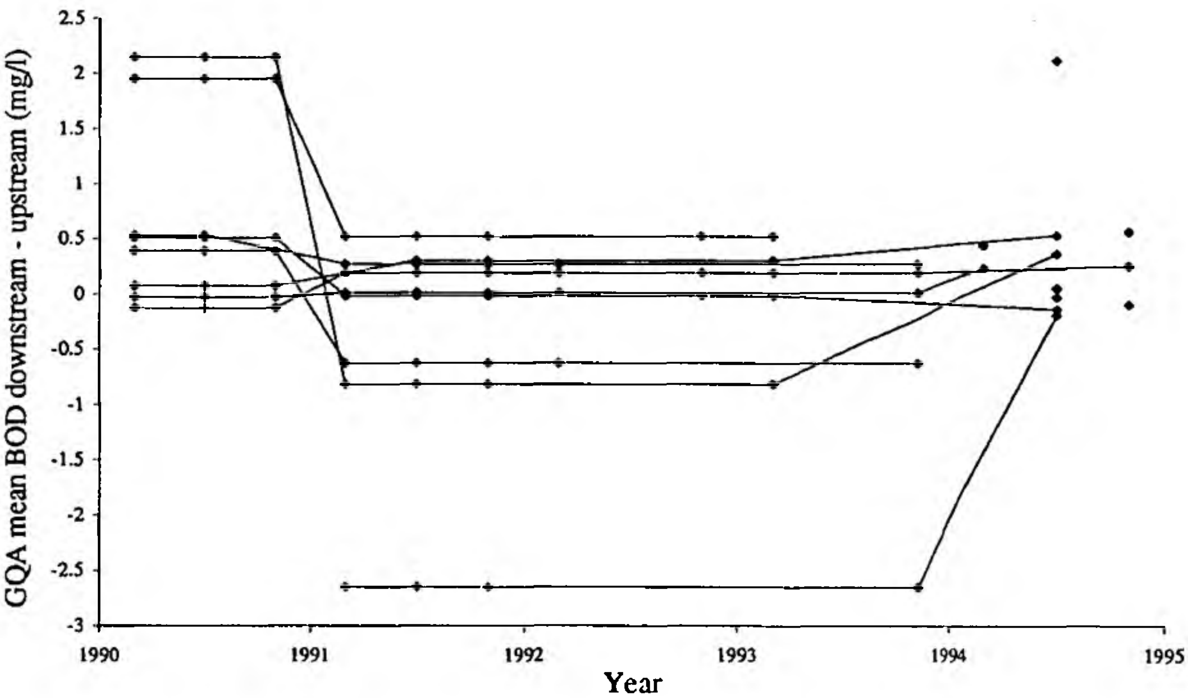

Figure 5.28 GQA BOD change plotted against time

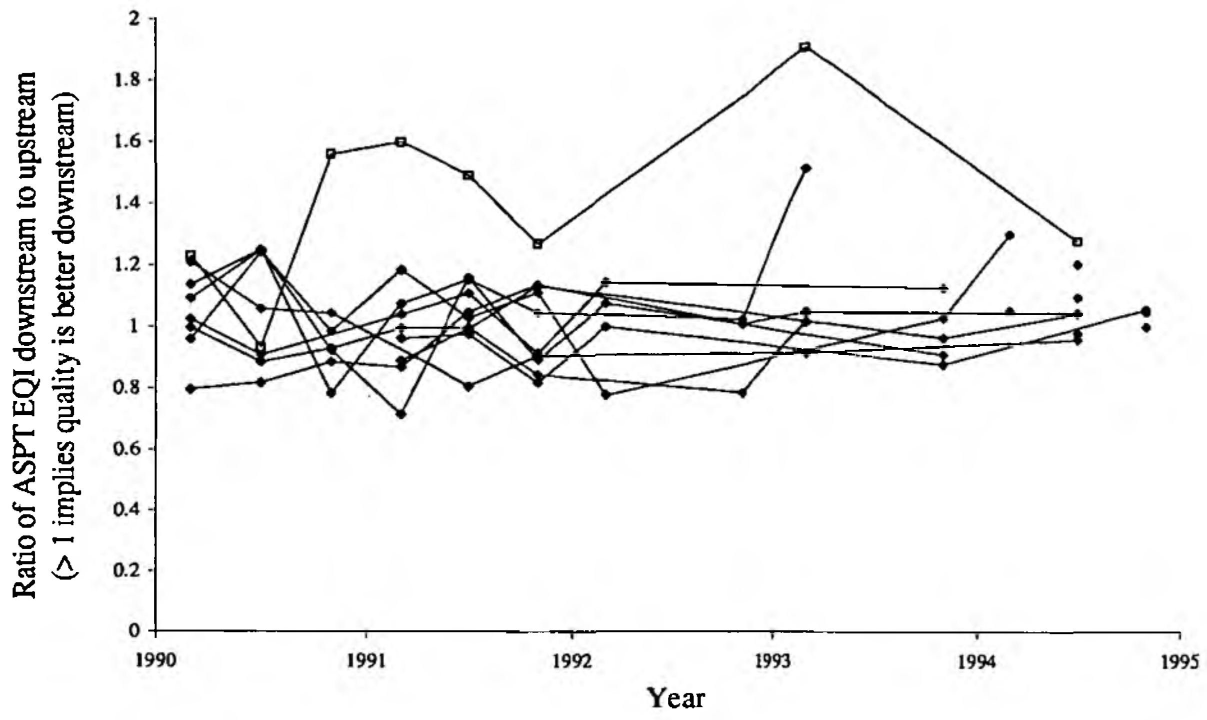

Figure 5.29 Ratio of ASPT EQI plotted against time

6. IDENTIFICATION OF RELATIONSHIPS BETWEEN EFFLUENT QUALITY AND BIOLOGICAL QUALITY

Interest lies in the relationship between the effluent quality of discharges and the changes in biological quality downstream of discharges. In order to optimally define this relationship, 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 are:

- theoretical increases to receiving water concentrations (BOD, ammonia and suspended solids),
- season (as defined in Section 4.2),
- 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.

6.1 Principal coordinates analysis

Principal coordinate 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 down 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.

The intention here is to reduce the biological change variables to a single measure and regress this single variable on the explanatory variables. A principal coordinates analysis of the changes-in-biological-quality variables produces encouraging results. Table 6.1 shows how much the information contained in the three biological variables can be explained by each of the principal coordinates. The first principal coordinate explains 82% of the variability by itself, and the first two coordinates explain 92.5%. 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 6.1 Percentage variation contained within the principal coordinates of the biological changes variables

Principal coordinate		
1st	2nd	3rd
82%	10.5%	7.5%

The transformation of the variables for the first coordinate turns out to be, approximately, the average of the standardised EQIs and LQI differences. In other words, a weighted average of the three variables with the weightings inversely proportional to the standard deviations of the variables.

Before regressing the first principal coordinate on the set of explanatory variables, it is important to reduce the size of the explanatory variable set. This is because there are variables within this set whose values are closely linked. Principal coordinate analyses were done on the upstream biology variables and the RIVPACS predictions, in each of which 87% of the variation was explained by the first coordinate, so both these groups of determinands can be reduced to one measure each. The first coordinate for the GQA means explains 61% of the total variation and the 2nd explains 26%. Therefore the GQA means were replaced by their first two coordinates.

In the case of the theoretical increases due to the effluent, BOD and ammonia were analysed using principal coordinates. Here, 98% of the variation was explained by the first coordinate; the transformation for which is a weighted average, with weightings inversely proportional to the standard deviation of the variables as per the biological-change variables. The increase in suspended solids was included unchanged as it is not too closely correlated with the other two effluent variables.

Table 6.2 Analysis of variance table for the regression of biological change on effluent quality

Source of variation	d.f.	s.s.	m.s.	v.r.
Theoretical increase due to effluent	1	0.09	0.0945	0.11
STW	31	54.1	1.744	1.6
Year	3	2.05	0.6846	0.77
Season	2	1	0.4986	0.56
Biology upstream	1	59.4	59.414	41.09
GQA upstream	2	0.01	0.0059	0.01
RIVPACS prediction	1	0.96	0.9604	1.07
IncSS	1	0.49	0.4948	0.55
Residual	106	130.4	1.23	
Total	147	364.9	2.483	

In the analysis of variance table (Table 6.2) for the regression, the sums of squares for each variable corresponds to the increase in residual sums of squares when removing that effect from the model. Since the design is not balanced the sums of squares do not add up to the total. STW, year and season are fitted as factors. A backward selection procedure was used and only the upstream biological variable (principal coordinate) and the STW factor were significant and together explained only 50% of the variation. Inclusion of the upstream biological variable, moreover, might have been in part an artefact due to the inbuilt negative correlation between biological change (i.e. downstream-upstream) and upstream. With all the x-variables in the model only 64% of the variation is explained.

This indicates that, as was shown in the initial plots, there is little or no discernible relationship between the effluent quality and the changes in biology.

6.2 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 pairs 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 analysis was done using the three biological-change variables as the set of y-variates and the full set of explanatory variables were used as the set of x-variates, 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 factor with 32 levels and the canonical correlation cannot cope with this since it would require the inclusion of 31 extra variables. The results of the analysis are shown in Table 6.3. The first canonical variables have a correlation of 0.83 which is equivalent to explaining about 68% of the variability between those two variables. This is a reasonably good correlation and of about the same order of goodness as the regression in the previous Section.

Table 6.3 Results of canonical correlation analysis

Canonical Variate	Canonical Correlation	% of the total Correlation
1st	0.83	37%
2nd	0.73	33%
3rd	0.68	30%

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 weighting 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 6.4.

To see how this works, suppose we increase the 'theoretical increases in BOD and ammonia' term in the model by 1 standard unit. Note that, because the weightings are inversely proportional to the standard deviations of the variables, this effectively amounts to an increase of one standard deviation (s.d.(theoretical increase)). Given this large increase, then the log BMWP EQI ratio would drop by 0.0656×1.0, or in other words the BMWP EQI would decrease by about 9% in real terms. This is a small change, implying that the relationship is weak. The other two coefficients in Table 6.4 indicate that a similar change would be expected for LQI and a still smaller change for ASPT.

Table 6.4 Coefficients of the variables in relationships with the biological-change variables as estimated by the canonical correlation

Explanatory variable	Log(BMWP EQI ratio)	Log(ASPT EQI ratio)	Difference in LQIs
Weighted average of the theoretical increases in BOD and ammonia	-0.0656	-0.0260	-0.0772
Increase due to effluent suspended solids	-0.0874	-0.0045	-0.1389
GQA mean BOD upstream	-0.0540	0.0168	0.8034
GQA mean ammonia upstream	0.0011	0.0023	0.0181
GQA mean dissolved oxygen	0.0036	-0.0013	-0.0093
Predicted BMWP upstream	0.1193	0.1942	0.5507
Predicted ASPT upstream	-0.0077	0.0013	0.0137
Predicted BMWP downstream	0.2119	-0.1414	-0.8602
Predicted ASPT downstream	-0.0076	0.0003	-0.0010
Observed BMWP upstream	-0.2860	-0.2133	0.1445
Observed ASPT upstream	0.0386	-0.0003	-0.7792
Summer	0.0083	0.0069	-0.1430
Winter	0.0423	-0.0314	-0.0913
Year 1991	0.0867	0.0176	0.4229
Year 1992	-0.0606	-0.3610	-7.2822
Year 1993	-0.0198	0.0096	0.6516
Year 1994	0.1158	0.0333	0.6064

7. CONCLUSIONS

Despite the use of a number of different statistical approaches, it has not proved possible to identify convincing relationships between STW effluent quality and receiving water biological quality from the available data.

Such a relationship may in fact exist but the available data was not adequate for identifying it. A number of features of the data point to this possibility:

- the final dataset was relatively small (29 STWs),
- biological data tended to show high variability within sites,
- few of the STWs were actually having much impact on river chemistry, nor, by implication, on biological quality.

Development of a larger database, in particular covering a greater range of STW impact on river chemistry, would increase the likelihood of identifying relationships between effluent quality and biology, assuming these do exist.

8. RECOMMENDATIONS

It is not worth pursuing the study further with the existing dataset, since the lack of convincing relationships between effluent quality and biology precludes the development of a protocol for using biology to assess effluent impact.

NRA Anglian Region should consider whether it wishes to pursue the project further through the augmentation of the dataset. Possible measures to achieve this are:

- careful audit of the existing dataset to ensure no data is missing, for example through mis-matched codes in the original databases,

- addition of more STWs from Anglian Region, particularly from other Areas, that provide a greater range of receiving water impact,

- addition of sites from other NRA Regions.

9. TARGETS AND TIMESCALES

The first stage of the project has been completed. There was a delay of several months in collation of all the data required, although the final delay in completion of Stage 1, and production of this report, was only one month.

The failure to find relationships between effluent quality and biology clearly has fundamental implications for the future course of the project. It would appear impossible, with the present data availability, to proceed with Stage 2. In light of this, the future of the project needs to be reviewed, and Section 8 gives some possible options for augmenting the data if this is considered worthwhile.

REFERENCES

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