# STATISTICAL MODELLING OF SEDIMENT CONCENTRATION

by

MAI PHUONG THOMPSON

B.Sc., The University of British Columbia, 1984

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE

in

THE FACULTY OF GRADUATE STUDIES
(The Department of Statistics)

We accept this thesis as conforming
to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA
August, 1987

4 6

Department of Statistics

The University of British Columbia
1956 Main Mall
Vancouver, Canada
V6T 1Y3

Date September 3, 1987

# ABSTRACT

One technique that is commonly used to replace the costly daily sampling of sediment concentration in assessing sediment discharge is the "rating curve" technique. This technique relies on the form of the relationship between sediment concentration and water discharge to estimate long-term sediment loads.

In this study, a regression/time-series approach to modelling the relationship between sediment concentration and water discharge is developed. The model comprises of a linear regression of the natural logarithm of sediment concentration on the natural logarithm of water discharge and an autoregressive time-series of order one or two for the errors of the regression equation. The main inferences from the model are the estimation of annual sediment loads and the calculation of their standard errors. Bias correction factors for the bias resulted from the inverse transformation of the natural logarithm of sediment concentration are studied. The accuracy of the load estimates is checked by comparing them to figures published by Water Survey of Canada.

# TABLE OF CONTENTS

# LIST OF TABLES

TABLE

## LIST OF FIGURES

FIGURE

# ACKNOWLEDGEMENT

I would like to thank Doctor H. Joe for his guidance and assistance in producing this thesis. I am indebted to Doctor K. Knight for his careful reading and helpful criticisms of this thesis. I would like to express my gratitude to Doctor M. Church, who kindly offered advice on the hydrological aspects and carefully read this thesis.

This thesis is funded by Canada Department of Supply and Services for the Sediment Survey Section, Water Survey of Canada, Water Resources Branch, Inland Waters Directorate, Environment Canada, by a contract to Doctor M. Church, Geography Department, U.B.C.

# CHAPTER 1

# INTRODUCTION

This thesis consists of a statistical study and modelling of sediment transport observations. The basic issue is the form of the relationship between streamflow and the concentration of suspended sediment. Important inferences from the model includes the estimation of long term suspended sediment loads and the calculation of their standard errors.

Information on suspended sediment load has many uses, including:

1. design and operation of reservoirs,
2. design of canals, diversion works and drainage ditches,
3. prediction, evaluation and control of deposition of natural channels,
4. evaluation of erosion control measures,
5. design of treatment facilities for water supply,
6. improvement of fish habitat.

More details on the purposes of estimating suspended sediment load can be found in the consulting report by Church *et al.* (1985).

Sediment data come from sediment and hydrometric stations. In 1985, there were approximately 185 (M. Cashman, Sediment Survey of Canada, personal communication, 1987) active suspended sediment stations in Canada for which daily, monthly and yearly suspended sediment concentrations and corresponding suspended loads were being measured. All these stations were operated in conjunction with hydrometric stations which generally predate the corresponding sediment stations by a number of years. The sediment records cover periods from 1 to a maximum of 23 years while corresponding discharge records cover over 72 years. There are presently over 3000 (Kellerhals *et al.* (1974)) operating hydrometric stations in Canada, thus exceeding the number of sediment of sediment stations by an order of magnitude.

There are many similarities between the operation of a sediment station and that of a hydrometric station. However, the hydrometric station costs much less to operate and is a prerequisite for the suspended sediment station. The cost-differential is due to the many more frequent visits, and the laboratory analysis of samples that are required in the operation of the suspended sediment stations.

Because of rapidly increasing costs, it is becoming more difficult to justify the continued operation of many of the daily sediment sampling stations on streams and rivers. In these circumstances, use of a relationship between suspended sediment transport and water discharge can reduce the need to sample daily sediment concentrations and hence reduce costs. Hence, the purpose of the thesis is to explore model structure and performance for inferring sediment load from discharge data.

Section 1 of this chapter gives a brief outline of the thesis. Section 2 provides a description of the data used in this study and Section 3 presents the definitions of relevant hydrological terms.

## 1.1 OUTLINE OF STUDY

This thesis is organized as follows. A review of some of the studies relevant to the present project is presented in Chapter 2. Included are notices of the statistical shortcomings of previous research papers.

In Chapter 3, the relationship between the mean daily sediment concentration, $C$, and the mean daily discharge, $Q$, is investigated. (Mean daily values are studied because they correspond with the resolution of normally available flow data.) Modification of the model

$$\ln C_t = \beta_0 + \beta_1 \ln Q_t + \epsilon_t$$

where $\{\epsilon_t\}$ are random errors, provides an adequate fit to the data. Various discharge-related variables are examined as possible additional predictors. The final model is fitted using least-squares regression. An analysis of the residuals shows that the model errors are serially correlated, so $\{\epsilon_t\}$ is modelled as a first-order autoregressive process.

The model is used for estimating the annual sediment loads of a river together with the standard errors. Detailed calculations are given at the end of Chapter 3.

Finally, in Chapter 4, summaries of estimates are given and the estimated annual loads are compared with those calculated by the Water Survey of Canada by direct integration from frequent observations.

## 1.2 DESCRIPTION OF DATA

### a) Sediment sampling stations

Four sediment stations are considered for this study. Two of the stations are located on the Fraser river, one at Mission (08MH024) and the other at Hansard (08KA004). Of the remaining two, one is located on the Oldman River, near Brocket (05AA024), in Alberta and the other is located on the Big Creek, near Walsingham (02GC007), in Ontario. The codes in parentheses are station reference codes used by the Water Survey of Canada.

Different geographical sites are picked for the purpose of checking to see if a general modelling approach remains valid in a variety of geographical locations with different characteristics.

**b) Suspended sediment concentration and load**

The equipment, methods and procedures used by the Water Survey of Canada in obtaining suspended sediment records have been described by Stichling (1969, 1973). The published data consist of average daily values for suspended sediment load, $L$, in tonnes per day, and suspended sediment concentration, $C$, in mg/l.

It should be noted that instantaneous concentration, $CI$, is the parameter actually determined in the field by a sampling program. The rivers are sampled at irregular intervals, ranging from fractions of a day up to one month, depending on the transport rate and its variability.

A continuous concentration graph is obtained by interpolating between the observed $CI$ values. This interpolation relies mostly on the stage records of the hydrometric station that is always associated with the sediment stations, but many other factors are also considered. An average concentration, $C$, is determined for each 24-

hour interval and combined with the mean daily discharge, $Q$, in cubic metres per second, to give the daily transport rate according to the following equation

$$L \text{ (tonnes per day)} = 0.0864\ Q\ (\text{m}^3/\text{s})\ C\ (\text{mg/l})$$

Ideally $CI$ should be used because $C$ is based on the stage record and therefore also the discharge. However, for the present sites, $CI$ and the corresponding discharge $QI$ were not readily available. In order to minimize the spurious effects of discharge, all those $C$ values for days on which no samples had been collected, were omitted from the analysis.

### c) Discharge

The basic data used are the mean daily flows as published by the Water Survey of Canada.

## 1.3 VARIABLE NAMES AND DESCRIPTIONS

*Sediment* is non-aqueous material transported by, suspended in or deposited by a flowing stream.

*Suspended sediment or suspended load* is solid material that moves in suspension in water, either as a colloid or through the influence of the upward component of turbulent currents.

*Suspended sediment concentration* is the ratio of the weight of dry solids in a water-sediment mixture to the volume of the mixture.

*Instantaneous concentration* is the above ratio of a sample taken instantaneously at a fixed point in a river cross section.

Some relevant variable names and their descriptions are given below. These notations are used in Chapters 2 and 3.

| | |
|---|---|
| $C$ | Mean daily concentration of sediment. |
| $Q$ | Mean daily discharge. |
| $CI$ | Instantaneous concentration of sediment. |
| $QI$ | Instantaneous discharge. |
| $\overline{C}_m$ | Mean concentration of sediment for month $m$. |
| $\overline{Q}_m$ | Mean discharge for month $m$. |
| $\overline{Q}_r$ | Mean discharge for the period of record. |
| $\overline{Q}_{pi}$ | Mean discharge for interval between samples. |
| $\ln C$ | Natural logarithm of the mean daily concentration of sediment. |
| $\ln Q$ | Natural logarithm of the mean daily discharge. |
| $\ln Q_i$ | Natural logarithm of the mean daily discharge for day $i$. |
| $\ln Q_{-i}$ | Natural logarithm of the lag i mean daily discharge. |
| $L$ | Total load for the period in context. |
| $K$ | Correction factor to take account of period of record. |
| $n$ | Number of samples. |

# CHAPTER 2

# REVIEW OF SOME PREVIOUS WORK ON RATING CURVES

Rating curves for sediment concentration based upon discharge have many uses, but their most extensive use is in estimating the sediment yield. Such rating curves are produced by fitting a function to the scatter plot of the sediment concentration transformation, which may be a daily, monthly or yearly average, against the corresponding discharge transformation. Commonly, log-transforms are used. Since the introduction of the rating curve by Campbell and Bauder (1940), much progress has been done.

A brief chronological review of some of the more recent work on rating curves and sediment load estimation is presented in Sections 2.1 to 2.4. Included are studies by Kellerhals, Abrahams and Von Gaza (1974), Walling and Webb (1981), Smillie and Koch (1984) and Church, Kellerhals and Ward (1985). Kellerhals *et al.* (1974) and Church *et al.* (1985) are mainly concerned with the development of a model for the sediment concentration, which may be a daily, monthly or yearly average, based on the corresponding discharge and some discharge-related variables. Walling and Webb (1981) assessed the performances of six frequently used methods of estimating long term sediment load and the reliability of load estimates obtained using rating curve technique. Smillie and Koch (1984) derived a bias correction factor, based on a normality assumption, for estimating the sediment concentration in the usual logarithmic model employed in all of the other studies.

## 2.1 KELLERHALS ET AL.'S (1974) CONSULTING REPORT

Using only the data of the high-flow period of five Western Canadian stations, Kellerhals *et al.* tested the effectiveness of the six factors in Table 2.1 and their logarithmic, square and square root transformations, when appropriate, in predicting $C$ in a linear regression model of the form

$$\ln C = \beta_0 + \beta_1 \ln X_1 + \beta_2 X_2 + \ldots + + \beta_p X_p + \epsilon.$$

Through stepwise regression, they discovered that except for the station with the smallest drainage basin, $X_1$ is the dominant and sufficient factor. This agrees with the findings of prior workers such as Abraham (1969) and Church (1972).

To test the rating equations, Kellerhals *et al.* calculated the estimated load for the high-flow period of each year and found that these estimates were generally too low. Furthermore, for virtually every year on record, the highest concentration and the mean of the ten highest concentrations observed generally exceed the corresponding concentrations computed from the rating curves. The study was not carried any further due to the difficulties faced by the authors in dealing with serially correlated residuals.

The general underestimation found by Kellerhals *et al.* could be attributed to the model being too simplistic and most likely to the bias resulted from inverting from $\ln C$ to $C$, as will be discussed later. As for the underestimation of the highest concentration and the mean of the ten highest concentrations, it should be noted that rating curves predict means, not extremes.

## 2.2 WALLING AND WEBB (1981)

Walling and Webb used a continuous record of sediment concentration from the River Creedy in Devon, U.K. to assess the accuracy and precision of some commonly used methods of estimating long-term sediment load. The authors divided these methods into two groups: one involving the use of the rating curves and the other, involving more direct approaches of estimation, namely

1. Total load = $K(\sum_{i=1}^{n} CI_i/n)(\sum_{i=1}^{n} QI_i/n)$

2. Total load = $K\sum_{i=1}^{n} (CI_i)(QI_i)/n$

3. Total load = $K\overline{Q_r}\sum_{i=1}^{n} CI_i/n$

4. Total load = $K\overline{Q}_r\left(\sum_{i=1}^{n}(CI_i)(QI_i)/\sum_{i=1}^{n}(QI_i)\right)$

5. Total load = $K\sum_{i=1}^{n}(CI_i)\overline{Q}_{pi}$

6. Total load = $K\sum_{m=1}^{12}\overline{C}_m\overline{Q}_m$

The variables are defined in Section 1.3.

To examine the accuracy and precision of the methods in the latter group, Walling and Webb applied the methods to replicate sets of data representing systematic sampling at one, two, four, seven, ten and fourteen day intervals. The authors argued that although methods 2 and 4 were by far the more accurate ones, other methods, particularly 1 and 3, yielded smaller variances and therefore were preferrable. However, it should be noted that the greater precision yielded by methods 1, 3, 5 and 6 is an artifact of their having averaged out some of the variability. Furthermore, these methods are justified only in the case where $C$ or $Q$ is approximately constant within the periods over which the averages were taken. Positive correlation between $C$ and $Q$ would result in underestimation of the total loads. Since for a positive correlation between $C$ and $Q$, $E(QC) > E(C)E(Q)$; hence on the average, the product between the average sediment concentration and the average flow is less than the average of the products of the two.

The great increase in variability of methods 2 and 4 as the sampling interval increases is due to the fact that in systematic sampling, longer interval means fewer points and poorer precision. It should also be noted that systematic sampling may give poor precision when periodicity is present (Cochran (1977)).

In evaluating the load estimates obtained using rating curves, four data sets from a seven year record, each containing fifty replicates representing four sampling strategies were assembled. The resultant rating curve relationships were applied to several frequently used load calculation procedures.

The sampling strategies employed to generate the replicate data sets included regular sampling at weekly intervals and coverage of storm events with additional random sampling when flows exceeded certain thresholds. Fifty replicate relationships of the form

$$C = aQ^b$$

where $Q$ is the instantaneous discharge at the time of sampling, were calculated using least squares regression. In addition, two of the data sets were subdivided into four stages: winter and rising stage, winter and falling stage, summer and rising stage, summer and falling stage; four separate relationships were developed for each stage. The results showed that accuracy and precision increased with the coverage of storm events. Most noteworthy is the discovery that the use of rating relationships subdivided according to season and stage tendency produced an increase in the accuracy of the estimate provided by a particular sampling strategy.

In general, all four strategies underestimated the total load for the seven year period. This could be the result of poor sampling and/or poor modelling in addition to the lack of a bias correction factor.

## 2.3 SMILLIE AND KOCH (1984) AND FERGUSON (1986)

For years, hydrologists have overlooked the bias in their estimation of the sediment loads or concentrations that resulted from simply transforming the fitted model

$$\ln \hat{C} = \hat{a} + \hat{b} \ln Q$$

to obtain

$$\hat{C} = \exp(\hat{a}) Q^{\hat{b}}$$

Smillie and Koch (1984) and Ferguson (1986) showed that if certain assumptions were satisfied, this bias could be easily corrected.

Smillie and Koch illustrated this biased behaviour for data on the Yampa River, near Maybell and the Little Snake River, at Lilly Park, in Northwestern Colorado, U.S.A., but found that for both rivers, the bias correction factor derived overcorrected the estimates of the annual sediment loads.

To further illustrate the tendency towards bias and to test the ability of the correction factor to improve a prediction, Smillie and Koch created a data set for which the following properties were satisfied:

1. A log-transformed linear relationship exists between the two variables.

2. Errors are normally distributed.

3. There is temporal stationarity in the system.

As in the case with the real data, the uncorrected model underpredicted the average of the dependent variable. On the other hand, the use of the bias correction factor improved the estimates considerably and no sign of overcorrection was detected.

The consistent overestimation noted earlier for the Yampa River and the Little Snake River could be the result of the following possibilities:

- The data collected were not representative.
- The model used was too simplistic.
- The normality assumption was not satisfied.

Smillie and Koch suggested that the serial correlation between the errors might have caused the overcorrection by the bias correction factor, but as shown in Appendix A.4, this should affect only the variance of the estimate and not the estimate itself.

Ferguson applied the bias correction factor to sixteen simulated and real data sets. For the nine simulation experiments, he obtained corrected estimates that lie within three percent of the true load. For the seven real data sets, the corrected estimates range from 91 to 104 percent of the true load. So, when the normality assumption is reasonable, the bias correction factor derived appears to be an adequate one.

## 2.4 CHURCH, KELLERHALS AND WARD'S (1985) CONSULTING REPORT

As in Walling and Webb (1981), Church, Kellerhals and Ward divided long-term sediment load estimation into two groups: one involving the use of rating curves and the other involving more direct approaches.

In the latter group, the authors divided the methods into three main approaches

1. Total load = $K \sum_{i=1}^{n} C_i Q_i / n$

2. Total load = $K(\sum_{i=1}^{n} C_i/n)(\sum_{j=1}^{m} Q_j/m)$

3. Total load = $K \sum_{m=1}^{12} C_m Q_m$

They dismissed the second and third approaches for reasons similar to those mentioned in Section 2.2 and recommended the first approach, it being the only "unbiased" one. It should be noted that the first approach is not necessarily unbiased if sampling is not random. The second approach is valid only if $Q$ is of no use in predicting $C$. The third approach is useful where $C$ and $Q$ are approximately "independent" within the time periods over which they were averaged.

Turning to the methods involving rating curves, Church, Kellerhals and Ward used data from six Western Canadian stations to run the following regressions:

1. Daily concentration vs. daily flow for all sampled days in each year,

2. As in 1 but including only every second, every fifth, or every tenth sample,

3. Daily concentration vs. daily flow for all samples in one, two, three and five years combined,

4. Annual sediment load vs. annual water volume for complete years of record.

In the analysis 1, the data of a year were used to derive a rating equation and this equation was in turn used to estimate the year's total sediment load. The results were very close to those figures provided by the Water Survey of Canada for the stations with big drainage areas and were much less accurate for those with small drainage areas. This is in agreement with the findings of other workers (Kellerhals *et al.* (1974)). The estimates, however, were generally smaller than the ones published by the Water Survey of Canada. This general underestimation was corrected when the bias correction factor suggested by Smillie and Koch (1984) was used.

In the analysis 2, the results were comparable to those obtained in the analysis 1. The authors concluded that if long-term sediment yield was the main object of observation, twenty to thirty samples a year were easily sufficient.

There was no systematic change in the quality of the annual load estimates obtained using the rating curves derived from two, three or five years in the analysis 3 in comparison with those obtained in the analysis 1, where annual rating curves were used. This suggested that the annual relationship between water discharge and sediment concentration does not vary drastically from year to year so that using the rating curve developed from data of past years to estimate future annual loads from flow data is not out of line. This would, however, be subject to vagaries of the long-term behaviour of the contributing drainage basin.

The analysis 4 showed good results for stations with large drainage areas. The procedure, however, was not very useful for stations with small drainage areas. Church, Kellerhals and Ward went further by examining flow during restricted portions of the year for one of the stations with larger drainage areas. The combinations were: May–June, May–July, April–July, April–August, May–September, and April–

October. The authors discovered that all combinations yielded better $R^2$ than the annual flow. This suggested that the linear relationship assumed between the logarithm of the sediment load and that of the flow is more of a high-flow period relationship than an annual one. These $R^2$ values should not be used as a means of comparisons between the portions of the year as time periods over which the modelling should be done since the dependent variable in each case is different.

In the end, Church, Kellerhals and Ward suggested a new method, to be used in case of sparse data, called the shifting rating curve. This method suggested that all available data be used to compute the best fitting curve of the form $C = aQ^b$. Then assuming that $b$ is fixed, $a$ is recalculated for each sampled concentration of sediment as follows

$$a_i = C_i {Q_i}^{-b}$$

Linear interpolation was used to estimate the coefficient between samples. This scheme was tested on for two stations, one using every tenth sample and the other using every fifth sample. These same samples were employed to derive the yearly rating curves, and comparisons amongst the annual loads computed by this new method, by the fixed yearly rating curve and by the Water Survey of Canada were carried out. The results showed substantial improvement in the annual loads estimated by the new method over those derived from the single year, fixed rating curves method. However, this method does not have a solid foundation. Furthermore, an implicit assumption is that the data be of the form shown in Figure 2.1.

### 2.5 DISCUSSION

With the exception of Smillie and Koch (1984), all investigators had problems with underestimating the load, even with the bias correction, when assuming that the daily relationship between $\ln C$ and $\ln Q$ over the year or over the high-flow period can be described by the model

$$\ln C = \beta_0 + \beta_1 \ln Q + \epsilon \tag{2.1}$$

Kellerhals *et al.* (1974) tried including several other independent variables but the results showed that only for rivers with small drainage areas did the addition of mostly one variable increase the percent of explained variation in $\ln C$. Even then, there was no obvious second best factor.

Assumptions required for the use of the adopted statistical model were mentioned by Kellerhals *et al.*, but none of the papers reviewed in this chapter reported on whether the data satisfied the assumptions. Also, checks for the inadequacies of the models used were not mentioned in any of the papers. Some authors noted that the residuals were highly serially correlated but the problem was not pursued.

With a record of continuous data, Walling and Webb (1981) were able to create replicate sets of data and obtain the sample standard deviation of the load. With the correct procedures and a sampling interval of one day, they obtained at best a standard deviation of about ten percent of the load. None of the investigators attempted to calculate the standard error of the load estimate based on a model.

In this study (described in Chapter 3), the relationship between the mean daily sediment concentration and the mean daily discharge is also found to be well represented by the model (2.1). However, as Walling and Webb (1981) have discovered, the fit is much improved when the seasonal effect is taken into account. Here, the data are split into two parts: before and after the day of the peak flow, and a different relationship is fitted to each set.

Following the work of Kellerhals *et al.* (1974), several discharge-related variables are included in the the relationship mentioned in the preceding paragraph. Of these, only the inclusion of a certain lag of the mean daily discharge is found to increase the portion of explained variance in $\ln C$ by a significant amount, and even then, only for one of the relationships.

As noted by Kellerhals *et al.* (1974), Smillie and Koch (1984), and Church *et al.* (1985), the residuals are positively serially correlated. This is not unexpected as the observations are ordered in time, hence adjacent cases influence each other. Analysis of the residuals suggests that the errors be modelled using a first-order autoregressive model

$$\epsilon_t = \rho\epsilon_{t-1} + \eta_t,$$

where $\{\eta_t\}$ are independent and identically normally (i.i.d.) distributed random variables with mean zero and constant variance. The standard assumption in regression analysis that the errors, $\epsilon_t$ are uncorrelated no longer applies.

This violation of the uncorrelated errors assumption does not affect the bias correction factor for $\hat{C}$, as suggested by Smillie and Koch (1984). The bias correction factor remains $\exp(\sigma^2/2)$ in the case of normally distributed error. However, the calculation of the load variance is very sensitive to the assumption of uncorrelated errors, as shown in Appendix A.4.

Up to now, only the bias correction factor for the case of normally distributed errors has been mentioned in the rating curves literature. In practice, the normality assumption may not always hold. In Section 3.3 and Appendix A.2, the sensitivity of the bias correction factor to the normality assumption is studied. Also a nonparametric estimate of the bias correction factor is introduced.

# CHAPTER 3

# STATISTICAL ANALYSIS AND MODELLING

## 3.0 INTRODUCTION

The models used in this chapter, which are based on the exploratory data analysis described in Section 3.1, have the form

$$\ln C_t = F(\ln Q_t, \cdots) + \epsilon_t, \quad \epsilon_t = \rho\epsilon_{t-1} + \eta_t, \tag{3.1}$$

where $\{\eta_t\}$ are independent and identically distributed error terms. That is, $\ln C_t$ is some function of $\ln Q_t$ and some other variables, plus an error term, which is serially correlated in time. The model is detailed in Section 3.2. One important inference from the model is the estimation of the sediment load for the high-flow period (Section 3.3). This involves a bias correction factor from the inverse transformation of the first term in (3.1). Some theory behind the bias correction factor is presented in Appendices A.1, A.2 and A.3.

Sections 3.1 and 3.2 consist of a detailed analysis and modelling of the Mission data. The Mission data are the most complete and "densely" collected set of data among the four stations under study, and perhaps in Canada. For this reason, a model developed for Mission can be used as a guide for modelling other stations, in particular Hansard, Brocket and Walsingham. Section 3.4 summarizes the analysis of these other three data sets.

Most statistical analyses were carried out using S and SAS. S is especially useful for exploratory data analysis; the main reference for it is Becker and Chambers (1984).

### 3.1 EXPLORATORY DATA ANALYSIS FOR MISSION DATA

To study the relationship between $C$ and $Q$ in relation to time, the yearly plots of $C$ and $Q$ versus time are examined. As shown in Figure 3.1, the variation of $C$ and $Q$ by several orders of magnitude makes it very difficult to identify the relationship between the two. To facilitate this task, the plots of the natural logarithm of $C$, $\ln C$, vs. the natural logarithm of $Q$, $\ln Q$, are inspected. These plots, as presented in Figure 3.2, show that there exists a relationship between $\ln C$ and $\ln Q$ during the high-flow period, usually between March and October; but no clear pattern can be detected during the low-flow period between November and Febuary (due to few observations).

In addition, these plots indicate that the relationship between $\ln C$ and $\ln Q$ varies from the rising stage of the high-flow period to the falling stage of the same period. This observation is made obvious by the yearly plots of $\ln C$ vs. $\ln Q$, with each point identified by the month in which it was sampled. As shown in Figure 3.3, these plots reveal a systematic annual hysteresis. That is, the trend of $\ln C$ vs. $\ln Q$ is reasonably consistent between June and October and in March, but between April and May, departs toward substantially higher $\ln C$ for a given $\ln Q$. For the remaining months, there appears to be no trend between $\ln C$ and $\ln Q$.

Due to having few observations and a lack of a clear pattern between $\ln C$ and $\ln Q$ in the low-flow months, only data from the high-flow period, between March and October, are included in further analyses. This omission should not affect the estimation of the annual sediment load as the amount of sediment transport during these low-flow months is observed in plots such as Figure 3.1 to contribute very little to the year's total load. A sample average of the mean of daily sediment concentration multiplied by the total flow will suffice as an estimate of the total sediment transport for these low-flow months.

Also omitted from the data are all samples from 1980. As a result of the very unusual weather in that year. The data do not follow the general pattern exhibited by other years. For example, the plot of $\ln C$ vs. $\ln Q$ for this year is very different pattern.

### a. A basic model

There is no theory which determines the form of the sediment rating curve. For suspended sediment, nearly all investigators (those whose papers are reviewed in Chapter 2 and references therein) have used a power law

$$C_t = aQ_t{}^b \exp(\epsilon_t) \tag{3.2}$$

where $\{\epsilon_t\}$ are random errors. This relationship is no doubt inspired by the linear trend discovered when plotting $\ln C$ against $\ln Q$. That is, by applying the natural logarithm to both sides of equation (3.2), a linear model of the form

$$\ln C_t = \ln a + b \ln Q_t + \epsilon_t$$

is obtained. The method of linear regression can then be applied.

### b. The seasonal effect

Many investigators (e.g., Miller (1951), Sharma and Dickinson (1980), Walling and Webb (1981)) believe that time is an important factor in studying the relationship between $\ln C$ and $\ln Q$. As observed earlier, the trend of $\ln C$ vs. $\ln Q$ appeared to vary from the rising limb of the hydrograph to the falling limb of the hydrograph. To substantiate this conjecture the following two models

$$\ln C_t = \begin{cases} \beta_{0r} + \beta_{1r} \ln Q_t + \epsilon_t, & \text{day } t \text{ is before the peak flow} \\ \beta_{0f} + \beta_{1f} \ln Q_t + \epsilon_t, & \text{day } t \text{ is after the peak flow} \end{cases}$$

and

$$\ln C_t = \beta_0 + \beta_1 \ln Q_t + \epsilon_t$$

are fitted to the data and the results compared. The gain in the portion of explained variance, the adjusted $R^2$, of the two-part model is approximately 16 percent. This suggests that two separate sub-models are required: one for the pre-peak flow days, the rising stage; and the other for the post-peak flow days, the falling stage.

### c. Other factors

The lags of the daily discharge are considered to be influential factors. It is generally observed that the sediment transport is dependent on not only the momentary discharge but also its recent history (Bogen (1980)).

To investigate the above observation, the natural logarithm of the lag 1 to lag 10 of the daily flow were examined. The scatter plots of the lags vs. $\ln C$ all exhibit a linear trend, although increasing variability is observed with the larger lags. An algorithm for selecting the "best" subsets of predictor variables in regression (*LEAPS* in S), is applied to the lags along with $\ln Q$. The results show that $\ln Q$ is the most dominant factor, explaining 59 percent of the variability in $\ln C$. The addition of the lag 10 discharge, $\ln Q_{-10}$, is accompanied by an increase of about 21 percent of the portion of explained variance in $\ln C$. (It should be noted that the addition of lag 7, lag 8 or lag 9 yielded comparable increase in the portion of explained variance as that of lag 10.) After these two variables have been introduced, further gain in the adjusted $R^2$ is minor. Thus, $\ln Q$ and $\ln Q_{-10}$ are used as predictors for $\ln C$.

According to Abraham (1969), Walling and Teed (1971) and Church (1972), other factors believed to influence the sediment transport include the daily rate of change of discharge ($rate$), the number of days since the last peak flow of the year ($dslp$) and the previous maximum discharge of the year ($prMadis$). These variables are thought to be related to the amounts of sediments available for transport. These variables along with $\ln Q_{-10}$ and $\ln Q$ will now be examined for the rising stage and the falling stage separately.

For the rising stage, all scatter plots (not given in the thesis) of the predictor variables vs. $\ln C$ , except the $dslp$ vs. $\ln C$ and $rate$ vs. $\ln C$ plots, exhibit some linear trend. The plot of $rate$ vs. $\ln C$ shows that the magnitude of $rate$, $abs(rate)$, is a better variable to use. Next, the best subsets algorithm is used to pick out the important factors. The results are summarized in Table 3.1 with $p$ as the number of parameters (the number of predictors plus the intercept).

With the adjusted $R^2$ serving as the criterion, the model with $\ln Q$ and $\ln Q_{-10}$ is the best model. The model with $\ln Q$ alone explains only 66 percent of the variability in $\ln C$. By including $\ln Q_{-10}$, the portion of explained variance in $\ln C$ is increased to 84 percent, which is the highest for any model with just two predictors. With $\ln Q$ and $\ln Q_{-10}$ already in the model, further gain in the adjusted $R^2$ is insignificant.

The Mallow's $C_p$ criterion (Draper and Smith (1981)) is also considered. The Mallow's $C_p$ statistic for a $p$-parameter model is defined by

$$C_p = RSS_p + 2p - n,$$

where $RSS_p = \sum(\ln C_t - \widehat{\ln C_t})^2$ and $n$ is total number of observations. Mallows (1973) suggests that good models will have negative or small $C_p - p$. Although the results do not appear to favour the above model, the ranking is the same as that provided by the adjusted $R^2$ criterion.

For the falling stage, the plots of the predictor variables vs. $\ln C$ are examined. Except for the plot of *prMadis* (which is now a constant, ie., the maximum discharge for the year) vs. $\ln C$, all scatter plots follow a definite linear trend. Again, the magnitude of *rate* is used instead of *rate* itself. The results obtained from running the best subsets algorithm are presented in the Table 3.2.

According to the adjusted $R^2$ criterion, $\ln Q$ is the only variable needed. The model with just $\ln Q$ yields an adjusted $R^2$ of 81 percent. Only a maximum of about 1 percent can be gained by adding more variables in the model.

Once more, the ranking based on the Mallow's $C_p$ criterion is very close to the ranking provided by the adjusted $R^2$ criterion.

In choosing the "best" subset model, it should be kept in mind that although criteria are provided to aid the investigator in making the selection, a model with a simpler form that does not give up much of the criteria-based performance is often preferred over more complex models. In keeping with this rule, the models mentioned above are preferred to their competitors.

### d. The "dog leg" effect

Next, a "dog leg" effect is considered; that is, the upper segment of the plot of $\ln C$ vs. $\ln Q$ has a tendency to flatten off or even approach horizontal. This can be explained by the fact that above a certain limit, sediment concentrations do not continue to increase with discharge at a uniform rate but may even remain constant when the supply potential of the catchment has been fully achieved (Gregory and Walling (1973)).

This effect can be seen in Figures 3.4a and b, where the plots of $\ln Q$ vs. $\ln C$ and $\ln C$ vs. $\ln Q_{-10}$ for the rising stage both follow a curved line. The curved lines running through the data are the nonparametric smooths proposed by Cleveland (1975). In S, these can be obtained through the function *LOWESS*; for the figures, the smoothing factor used is 0.2.

When the model

$$\ln C_t = \beta_0 + \beta_1 \ln Q_t + \beta_2 \ln Q_{-10t} + \epsilon_t$$

is fitted to the rising stage data, the plots of the Studentized residuals vs. the predicted values and the Studentized residuals vs. $\ln Q$ both betray signs of nonlinearity as shown in Figures 3.5a and b. The plot of the Studentized residuals vs. $\ln Q_{-10}$, on the other hand shows no indication of nonlinearity. This suggests that the model failure is more a function of $\ln Q$ than it is of $\ln Q_{-10}$. Thus some transformation of $\ln Q$ is required.

Careful examination of the $\ln C$ vs. $\ln Q$ plot reveals that only a slight transformation of $\ln Q$ is necessary. Figure 3.4a shows that the relationship between $\ln C$ and $\ln Q$ is approximately piecewise linear with a corner at $\ln Q = 8$. Thus the data for the rising stage can be fitted using two different lines joined at $\ln Q = 8$, namely

$$\ln C_t = \begin{cases} \beta_0 + \beta_1 \ln Q_t + \beta_2 \ln Q_{-10t} + \epsilon_t, & \text{for } \ln Q_t > 8 \\ \beta_0' + \beta_1' \ln Q_t + \beta_2 \ln Q_{-10t} + \epsilon_t & \text{for } \ln Q_t \leq 8 \end{cases}$$

subject to continuity at $\ln Q_t = 8$. This is equivalent to fitting the following linear model

$$\ln C_t = \beta_0 + \beta_1 X1_t + \beta_2 X2_t + \beta_3 \ln Q_{-10t} + \epsilon_t$$

where $X1 = (\ln Q - 8) \times (1 - d)$ and $X2 = (\ln Q \times d) + ((1 - d) \times 8)$, with $d = 1$ if $\ln Q > 8$ and $d = 0$ otherwise.

**e. Outliers**

The Studentized residuals (Weisberg (1980)) vs. the predicted values plots for the rising stage and the falling stage both display a few candidates for outliers. Because the least squares estimators can be very sensitive to outliers, alternative estimation methods, generally called robust regression methods, are used for comparison. These methods basically give lower weights to observations which seem inconsistent with the rest of the data, assuming that the fitted model is true. So, instead of minimizing $\sum(\ln C_i - \widehat{\ln C_i})^2$, the function: $\sum \rho([\ln C_i - \widehat{\ln C_i}]/s)$, is minimized. $\rho$ is a function that penalizes outliers (less severely than the least squares procedure) and $s$ is a fixed scale factor. Two such well-known methods are by Huber (1972) and Andrews (1974). The loss functions for these two are as follow:

$$\text{Huber}: \quad \rho(x) = \begin{cases} A^2[1 - cos((x/A)], & |x| \leq \pi; \\ 2A^2, & \text{otherwise.} \end{cases}$$

$$\text{Andrews}: \quad \rho(x) = \begin{cases} x^2/2, & |x| \leq K; \\ K|x| - K^2/2, & \text{otherwise.} \end{cases}$$

Displayed in Tables 3.6 and 3.7 are the results obtained using the Huber and Andrews criteria (with the constants $A$ and $K$ being 1.339 and 1.345 respectively) and the sum of absolute deviations criterion. These were obtained using the functions *RREG* and *L1FIT* in S. The estimates are found to be very close to those obtained using least squares. Thus, it is concluded that the possible outliers observed are not influential in the estimation of the parameters.

**f. Model for the errors**

So far, all the regression methods have been used in an exploratory analysis sense, because the assumption of independent and identically normally distributed errors is not necessarily valid. This assumption is now checked by studying the residuals plots.

The normality assumption is checked by plotting the normal probability plot of the Studentized residuals. The plot for the rising stage is presented in Figure 3.6a and the plot for the falling stage is presented in Figure 3.6b. Both these plots indicate slightly longer tails than those of the normal distribution. Otherwise, neither suggests any sign of serious violation of the normality assumption.

Because the data in this study form a time series, adjacent errors would be expected to be serially correlated. Residuals of both the rising stage and the falling stage were joined together and daily runs of residuals were plotted against time. The plots show that the successive residuals tend to be more alike than otherwise, indicating the presence of positive serial correlation. On a set of residuals for which lag 1 up to lag 6 exist, the plots of the residuals vs. their lags were examined. The plots exhibit a "lower left to upper right" tendency that is representative of positive serial correlation. The amount of scatter in the plots increases with the lag. The plots of the residuals vs. lag 1 and lag 2 are provided in Figures 3.7a and 3.7b.

An estimate of the lag i correlation is

$$\widehat{\rho(i)} = \frac{n_i^{-1} \sum e_t e_{t-i}}{n_0^{-1} \sum e_t^2}$$

where $n_i$ is the number $e_t$ for which $e_{t-i}$ also exists (ie., a measurement was taken) and $n_0$ is the length of the time series. The estimates of the lag 1 to lag 10 correlations are summarized in Table 3.3. These estimates are then plotted against their respective lags. The plot, as shown in Figure 3.8, indicates a decay of the correlation as the lag increases.

The partial autocorrelations (Box and Jenkins (1976)), $pac(i)$, are also estimated and the estimates are provided in Table 3.3. The partial autocorrelations form a set of statistical measurements, similar to autocorrelations, that reveal how time series values are related to each other at specified lags. The plot of the partial autocorrelations vs. the lags is shown in Figure 3.9. This figure indicates that a first or second order autoregressive model might be appropriate because the $\widehat{pac(i)}$ for $i \geq 3$ are approximately zero, whereas $\widehat{pac(1)}$ and $\widehat{pac(2)}$ are considerably larger.

The Akaike's Information Criterion (Priestley (1981)) , $AIC$, is a method that can be used to select an appropriate order of an autoregressive model. This criterion picks the order $k$ for which

$$AIC(k) = n \ln \hat{\sigma}^2 + n \sum_{i=1}^{k} \ln(1 - \widehat{pac(i)}^2) + 2k,$$

where $\hat{\sigma}^2$ is the residual mean square from the regression model for $\ln C$, is minimized. However, it is also desirable to have as simple a model as possible. With this additional objective, the selected $k$ should have a small $AIC(k)$, and none of the $AIC(i)$, for $i > k$, should be much smaller than $AIC(k)$ (that is, $AIC(i)$ stabilizes for $i > k$).

The $AIC(i) - n \ln \hat{\sigma}^2$, for $i$ from one to ten, are listed in Table 3.4. Following the above guidelines, a first order autoregressive model is preferred over the second order autoregressive model. This choice is also supported by the adjusted $R^2$ criterion. The adjusted $R^2$'s obtained by fitting the models

$$\epsilon_t = \rho \epsilon_{t-1} + \eta_t$$

and

$$\epsilon_t = \rho_1 \epsilon_{t-1} + \rho_2 \epsilon_{t-2} + \eta_t$$

to the set of residuals for which lag 1 and lag 2 exist, are almost equal.

## 3.2 MODEL FOR MISSION DATA

### a. The proposed model

Summarizing the analysis in Section 3.1, the models proposed are

$$\ln C_t = \begin{cases} \beta_0 + \beta_1 \ln Q_t + \beta_2 \ln Q_{-10t} + \epsilon_t, & \text{for } \ln Q_t > 8 \\ \beta_0' + \beta_1' \ln Q_t + \beta_2 \ln Q_{-10t} + \epsilon_t, & \text{for } \ln Q_t \leq 8 \end{cases}$$

subject to continuity at $\ln Q_t = 8$, for the rising stage and

$$\ln C_t = \beta_0 + \beta_1 \ln Q_t + \epsilon_t$$

for the falling stage, where

1. $\ln Q_t$ is considered fixed and $\ln C_t$ is random.
2. $\{\epsilon_t\}$ is a sequence of normal random variables with mean zero and variance $\sigma^2$ such that $\epsilon_t = \rho\epsilon_{t-1} + \eta_1$, where $\{\eta_t\}$ are independent and identically normally distributed with mean zero and constant variance.

The parameters $\beta_0$, $\beta_0'$, $\beta_1$, $\beta_1'$, $\beta_2$, $\sigma$ and $\rho$ can be estimated using the maximum likelihood estimators derived in Appendix A.6. However, this method is not available in statistical packages and it may not be robust to the normality assumption. An alternative method that involves two stages is to estimate $\beta_0$, $\beta_0'$, $\beta_1$, $\beta_1'$, $\beta_2$ and $\sigma$ using least squares regression and to estimate $\rho$ using least squares regression on the residuals of the first regression. The estimates obtained by this latter method are presented in Table 3.5.

### b. Model adequacy and assumptions

Some model inadequacies are more serious than others. A small deviation from normality for the errors is among the less serious ones. As checked earlier, the normal probability plots of the Studentized residuals for the rising stage and the falling stage both look satisfactory.

A potentially more serious failure is the heterogeneity of error variance. Heterogeneity of variance can be detected by plotting the Studentized residuals against the predicted values and the independent variables. For the rising stage, the plots of the Studentized residuals vs. the predicted values and the independent variables, presented in Figures 3.10a - c, all display a slight degree of heterogeneity. However, since a main goal is to estimate the annual load of sediment transport and not the daily concentration of sediment, the slight nonconstancy of variance, which may or may not be a result of outliers, should hopefully not be a factor. The plots of the Studentized residuals vs. the predicted values and the independent variable for the falling stage are presented in Figures 3.11a - b. Neither of the plots shows any indication of heterogeneity of variance.

A much more serious failure is the nonlinearity of the model. To check for nonlinearity, the plots of the Studentized residuals vs. the predicted values are studied. As shown in Figures 3.10a and 3.11a, such plots for the rising stage and the falling stage are well behaved. The specified models are considered are adequate.

### 3.3 ESTIMATION OF THE ANNUAL SEDIMENT LOAD

In Section 3.2, a model for the natural logarithm of the mean daily sediment concentration was constructed. To obtain an estimate of the mean daily sediment concentration itself, the transformed scale prediction $\widehat{\ln C}$ is retransformed by exp(.).

However, because the transformation is nonlinear and because

$$E(C) = \exp(\beta_0)Q^{\beta_1}E(\exp(\epsilon)) \neq \exp(\beta_0)Q^{\beta_1}$$

$\exp(\widehat{\ln C})$ is a biased estimator (see Duan (1983) and Miller (1984) for example ). Still, if the errors are normally distributed, a major portion of the bias can be eliminated by applying the adjustment factor $\exp(\hat{\sigma}^2/2)$ of Smillie and Koch (1984) (see Appendix A.1).

In practice, departures from the normality assumption are not uncommon. Some frequent departures from normality is that the underlying distribution of errors is approximately be symmetric but nonnormal, possibly "more peaked" than the normal with "lighter tails", or "less peaked" than the normal with "heavier tails". A family of distributions that includes some of the distributions described above is the generalized Gaussian distribution

$$f(x) = const \ \exp(-|x/b|^p).$$

A study of the sensitivity of the bias correction factor to the change in $p$, for $1 \leq p \leq 2$ is presented in Appendix A.3.

The biasing factor $E(\exp(\epsilon))$ can also be estimated nonparametrically using

$$\sum_{i=1}^{n} \exp(e_i)/n$$

. This estimator is approximately unbiased and does not assume any distribution for the errors. A comparison between the estimates of this bias correction factor and the "normal" bias correction factor is presented in Table 4.5. The two are very close, suggesting that any deviation from normality assumption is slight. A theoretical relative efficiency comparison between the two bias correction factors, under the normality assumption, is provided in Appendix A.2. The efficiency $\exp(\hat{\sigma}^2/2)$ to the nonparametric estimate is greater than 1.

After obtaining an estimator for the mean daily sediment concentration, the problem of estimating the sediment transport of the high-flow period can be addressed as follow. First, linear interpolation between $\widehat{C_t}$'s and between $Q_t$'s is performed to complete the sediment hydrograph and the discharge hydrograph. Then an integration process detailed in Appendix A.4 is used to determine the daily sediment loads. These daily loads are then summed to produce the total sediment transport estimate of the high-flow period.

For the low-flow period, the total sediment transport can be estimated as described in the third paragraph of Section 3.1. At Mission, it is observed that the mean daily sediment concentration remains approximately constant between December and February, but is higher in November. Therefore the sediment transport is determined separately for these two periods.

The expectation and variance of the sediment load of the high-flow period are calculated in Appendix A.5. These are used to obtain standard error of the high-flow period load. Discussion of estimates of total annual loads and their standard errors is given in Chapter 4.

### 3.4 OTHER DATA SETS

In Sections 3.1 and 3.2, the Mission data have been carefully examined and a model has been developed for the natural logarithm of the sediment concentration. The mean daily discharge, the lag 10 of the mean daily discharge and the season are the factors found to be important in defining the model. The daily rate of change of discharge, the number of days since the last peak and the previous maximum discharge are not as useful; hence, they are not included in the analyses of the other stations.

Basically, the same analysis was carried out for each of the remaining stations. The only difference being the additional use of a nonparametric smoothing function developed by Friedman and Stuetzle (1982), called *SUPERSMOOTHER*, in identifying the "peak" that separates the rising from the falling stage. For Mission, the day of the peak flow is generally the dividing point between the two stages. For the other data sets, there are some years in which there is a period after the peak flow during which it is not clear whether the general trend of the hydrograph is up or down. The *SUPERSMOOTHER* is used in these cases to pick out the local or global peak, which is identified with the peak of the smooth line. It should also be noted that for the Walsingham data, the year is reset to start on August 1 and ends on July 31. This is so that the full high flow period occurs sometimes in the middle of the year, as for the west coast stations, and the same analysis can be carried out.

The models developed for Hansard, Brocket and Walsingham are listed in Table 3.8. The fitted models are listed in Table 3.5. It is found that a lag of the mean daily discharge of about 8 to 10 days is needed in the rising stage model for all the west coast stations. For Walsingham, the lag 1 of discharge is included in the falling stage model. The seasonal effect, as indicated in the preceding statements, is important at all four stations.

Robust regressions were performed to test the sensitivity of the models to outliers. The results are listed in Tables 3.6 and 3.7. The estimates are close for all stations, indicating that the possible outliers are not influencial in the estimation of the parameters.

The error model used for all four stations is a first-order autoregressive model. The *AIC* suggested a second-order autoregressive model for Walsingham. However, the standard errors of the load estimates are not very different one way or the other, thus, the simpler model is used.

The residual plots show that all four stations have a few possible outliers. However, no serious violations of the normality assumptions are noted in the normal probability plots. A slight degree of heterogeneity of variance is observed in some residual plots, but probably not serious enough to affect the estimation of the annual sediment load.

Finally, it should be noted that the models developed for Brocket and especially Walsingham are not as effective, in terms of portion of explained variance in $\ln C$, as the ones developed for Mission and Hansard. The same basic factors that explain between sixty and eighty percent of the variability in $\ln C$ at Mission, Hansard and Brocket only explain about forty five percent of the variability in $\ln C$ at Walsingham. This finding is in agreement with the findings of other workers (Kellerhals *et al.* (1974) and Church *et al.* (1985)), in that models that work well for stations on big rivers tend to do poorly for those on small rivers. The reason could be that in small rivers, the sediment transport is more strongly regulated by short-term events such as storms, whereas in large rivers, these events would not be individually significant factors.

# CHAPTER 4

# COMPARISON OF ESTIMATED ANNUAL LOAD WITH PUBLISHED DATA

One purpose of the models developed in Chapter 3 is to reduce the costly sample collection of the sediment concentration. A test of the usefulness of these models can therefore be made by computing the annual sediment transport of past years and comparing results with figures published by the Water Survey of Canada.

In this present study, the annual sediment transport is calculated for several years per station and tabulated in tables 4.1 - 4.4 along with the published data. A simple computer program was written to calculate these values and the standard errors. It is found that although the estimates obtained are generally close to the values provided by the Water Survey of Canada for all stations, the accuracy is more consistent at Mission and Hansard than the others.

As a further test, cross-validation estimates for the annual sediment transport are calculated for three years per station. That is, for each year, the year's sediment load is calculated using the model based on other years. The cross-validation test is especially important in this study as it will indicate whether or not future annual sediment loads can be well estimated using the model based on past years. As shown in Table 4.5, these estimates are as good as those obtained using the models based on data from all years. Hence, it is reasonable to assume that future annual loads can be as well estimated by the model based on past and present years as they can be by the models based on all years, including those for which the loads are estimated.

Finally, to assess the precision of the estimates produced by the developed models, the standard errors of the annual sediment loads are computed (see formulas in

Appendix A.4). The results listed in Tables 4.1 - 4.4 show that the estimates obtained for Mission and Hansard are much more precise than those obtained for Brocket and Walsingham. The standard errors calculated for Mission and Hansard are about 10 to 15 percent of the load estimates while the standard errors calculated for Walsingham are about 20 percent of the estimated loads and those for Brocket vary between 30 and 65 percent.

From the above calculations, it can be concluded that the proposed models for Mission and Hansard are very capable of replacing the sampling scheme presently employed by the Water Survey of Canada. The models for Brocket and Walsingham, on the other hand, are not as successful. Although the annual sediment transport can be reasonably well approximated using the proposed models, the precision of the estimates, in terms of standard error, is lower than ideal for Walsingham and unacceptably low for Brocket.

The above findings are not surprising. Other workers (Kellerhals *et al.* (1974), Church *et al.* (1985)) have found that sediment transport in small rivers is more unpredictable than in large rivers. The reason could be that in small rivers, the sediment transport is more strongly regulated by the local events such storms, whereas in large rivers, the effect of these events is somewhat "averaged" out. The sediment transport in large rivers is more strongly regulated by larger events such as snowmelt and seasonal yield of sediment from the land surface. This reasoning is supported by the improvements attained by some workers when additional information regarding rainfalls are included in the model (Church (1972)).

# References

[1] Abraham, C. E. (1969). Suspended sediment discharges in streams, *A. G. U. Golden Anniversary Meeting in Washington, D. C.*

[2] Andrews, D. F. (1974). A robust method for multiple linear regression, *Technometrics*, **16**, 523-531.

[3] Becker, R. A. and Chambers, J. M. (1984). *S an Interactive Environment for Data Analysis and Graphics*, Belmont: Wadsworth.

[4] Bickel, P. J. and Doksum, K. A. (1977). *Mathematical Statistics: Basic Ideas and Selected Topics*, San Francisco: Holden-Day.

[5] Bogen, J. (1980). The hysteresis effect of sediment transport systems, *Norsk geog. Tidsskr.*, **34**, 45-54.

[6] Box, G. E. P. and Jenkins, G. M. (1976). *Time Series Analysis: Forecasting and Control* (revised ed.), San Francisco: Holden-Day.

[7] Campbell, F. B. and Bauder, H. A. (1940). A rating-curve method for determining silt-discharge of streams, *American Geophysical Union Trans.*, **21**, 603-607.

[8] Church, M. (1972). Baffin Island sandurs: a study of Arctic fluvial processes, Geographical Survey of Canada, Department of Energy Mines and Resources.

[9] Church, M., Kellerhals, R. and Ward, P. R. B. (1985). Sediment in the Pacific and the Yukon Region: review and assessment, Report, Canada Department of Environment and Inland Waters Directorate.

[10] Cleveland, W. S. (1985). *The Elements of Graphing Data*, Monterey: Wadsworth.

[11] Cochran, W. G. (1977). *Sampling Techniques* (3rd ed.), New York: Wiley.

[12] Draper, N. and Smith, H. (1981). *Applied Regression Analysis (2nd ed.)*, New-York: Wiley.

[13] Duan, N. (1983). Smearing estimate: a nonparametric retransformation method, *J. Amer. Statist. Assoc.,* **78**, 605-610.

[14] Ferguson, R. I. (1986). River loads underestimated by rating curves, *Water Resources Research*, **22**, 74-76.

[15] Friedman, J. H. and Stuetzle, W. (1982). Smoothing of scatterplots, Department of Statistics, Stanford University, Tech. Report ORION006.

[16] Gregory, K. J. and Walling, D. E. (1973). *Drainage Basin Form and Process*, London: Edward Arnold.

[17] Hoff, C. J. (1983). *A practical Guide to Box-Jenkins Forecasting*, Belmont: Lifetime Learning Publications.

[18] Huber, P. J. (1972). Robust statistics: a review, *Ann. Math. Statist.,* **43**, 1041-1067.

[19] Huber, P. J. (1973). Robust regression: asymptotics, conjectures and Monte Carlo, *Ann. Statist.,* **1**, 799-821.

[20] Kellerhals, R., Abrahams, A. D. and von Gaza, H. (1974). Possibilities for using sediment rating curves in the Canadian Sediment Survey Program, Report, Canada Department of Environment, Alberta Department of Highways and Transport, Research Council of Canada and University of Alberta, Department of Civil Engineering, REH/74/1.

[21] Millard, S. P. and Guttorp, P. (1985). Hypothesis tests for regression models with autocorrelated errors. SIMS Tech. Report 106, Department of Statistics, University of British Columbia.

[22] Miller, C. R. (1952). Analysis of flow-duration, sediment rating curve method of computing sediment yield, Report, U. S. Department of Interior Bureau of Reclamation.

[23] Miller, D. M. (1984). Reducing transformation bias in curve fitting, *Amer. Statist.*, **38**, 124-126.

[24] Sharma, T. C. and Dickinson, W. T. (1980). System model of daily sediment yield, *Water Resources Research*, **16**, 501-506.

[25] Smillie, G. M. and Koch, R. W. (1984). Bias in hydrologic prediction using log-log regression model, *A. G. U. Fall Meeting.*

[26] Stichling, W. (1969). Instrumentation and techniques in sediment surveying, In: *Hydrology*, (Proc. Victoria, B. C. Symp., May 1969).

[27] Stichling, W. (1973). Sediment loads in Canadian rivers, In: *Fluvial processes and sedimentation*, (Proc. University of Alberta, May 1973).

[28] Walling, D. E. and Teed, A. (1971). A simple pumping sampler for research into suspended sediment transport in small catchments, *J. Hydrol.*, **13**, 325-337.

[29] Walling, D. E. and Webb, B. W. (1981). The reliability of suspended sediment load data, In: *Erosion and Sediment Transport Measurement*, (Proc. Florence Symp., June 1981), **IAHS 133**, 177-194.

[30] Water Survey of Canada, Environment Canada, Sediment Data for Canadian Rivers, Environment Canada, (Formely: Inland Waters Branch, Department of Energy, Mines and Resources, Ottawa, Canada), Yearly publications, 1966-1983.

[31] Weisberg, S. (1980). *Applied Linear Regression*, New York: Wiley.

# APPENDIX

### A.1 Bias correction factor for $\hat{C}$ with normal errors.

The estimator for the sediment concentration obtained by taking the inverse transformation of $\widehat{\ln C}$ of a model like (A1.1) below, is a biased one. On the assumption that the errors are normally distributed, a major portion of the bias can be removed by applying the correction factor derived here.

Suppose for simplicity that the true model for $\ln C$ is

$$\ln C = \beta_0 + \beta_1 \ln Q + \epsilon, \qquad (A1.1)$$

where $\epsilon$ is normally distributed with mean 0 and constant variance $\sigma^2$, and the fitted model is

$$\widehat{\ln C} = \hat{\beta}_0 + \hat{\beta}_1 \ln Q.$$

Inversely transforming the true and the fitted model yields

$$C = \exp(\beta_0) Q^{\beta_1} \exp(\epsilon) \quad \text{and} \quad \hat{C} = \exp(\hat{\beta}_0) Q^{\hat{\beta}_1}.$$

$\hat{C}$ is a biased estimator of $E(C) = \exp(\beta_0) Q^{\beta_1} E(\exp(\epsilon))$. The bias of $\exp(\hat{\beta}_0)$ for $\exp(\beta_0)$ and $Q^{\hat{\beta}_1}$ for $Q^{\beta_1}$ will be small since both $\hat{\beta}_0$ and $\hat{\beta}_1$ are based on a large number of observations. Hence, an approximately unbiased estimator of $E(C)$ can be obtained by multiplying $\hat{C}$ by an estimate of $E(\exp(\epsilon))$ (Duan (1983) and Miller (1984)).

If $\epsilon \sim N(0, \sigma^2)$, then $E(\exp(\epsilon)) = \exp(\sigma^2/2)$, by evaluating the moment generating function of $\epsilon$ at 1. So, assuming a normal error, an estimated correction term is

$\exp(\hat{\sigma}^2/2)$, where $\hat{\sigma}^2$ is the residual mean square obtained from the regression of $\ln C$ on $\ln Q$. Similarly, this extends to a more general regression model.

In this study, the correction is not done until after the load for the high-flow period has been estimated. Hence, the estimated correction term used is $\exp(\hat{\sigma_p}^2/2)$, where $\hat{\sigma_p}^2$ is the pooled residual mean square from the rising stage and the falling stage: $[(n_r - 1)\hat{\sigma_r}^2 + (n_f - 1)\hat{\sigma_f}^2]/(n_r + n_f - 2)$. Although it is preferrable to correct the individual $\widehat{\ln C_i}$, $\hat{\sigma_r}^2$ and $\hat{\sigma_f}^2$ are close enough so that it makes no difference.

### A.2 A comparison of two estimated bias correction terms.

As discussed in Appendix A.1, $\exp(\widehat{\ln C})$ is a biased estimator of $E(C)$. The biasing factor is $E(\exp(\epsilon))$. If the errors are normally distributed, this is equal to $\exp(\sigma^2/2)$ and an estimate of the biasing factor is $\exp(\hat{\sigma}^2/2)$, where $\hat{\sigma}^2$ is the residual mean square obtained from the least squares regression. For the data in this study, the normality assumption may be in doubt. A nonparametric estimate of the biasing factor is $n^{-1}\sum_{i=1}^{n}\exp(e_i)$, where $\{e_i\}$ are the residuals of the regression model.

In this appendix, the asymptotic relative efficiency (Bickel and Doksum (1977)) of the nonparametric estimate, when the errors are normal, is studied. Assume the errors are correlated as in (A6.1). By standard asymptotic theory, the Central Limit Theorem, and formulas for the moments and moment generating function of a bivariate normal distribution,

$$\sqrt{n}\Big(\exp(\hat{\sigma}^2/2) - \exp(\sigma^2/2)\Big) \xrightarrow{\text{d}} N\ (0,\ v_1{}^2),$$

where $v_1{}^2 = \exp(\sigma^2)\sigma^4(1+\rho^2)/2(1-\rho^2)$, and

$$\sqrt{n}\Big(\frac{1}{n}\sum_{i=1}^{n}(\exp(e_i)) - \exp(\sigma^2/2)\Big) \xrightarrow{\text{d}} N\ (0, v_2{}^2),$$

where ${v_2}^2 = \exp(\sigma^2)\Big[\Big(\exp(\sigma^2) - 1\Big) + 2\sum_{k=1}^{\infty}\Big(\exp(\sigma^2\rho^k) - 1\Big)\Big] = \exp(\sigma^2)\big(\sum_{k=1}^{\infty}(k!)^{-1}\sigma^{2k}[(1 + \rho^k)/(1 - \rho^k)]\big)$.

It can be seen that $v_1^2$ is less than $v_2^2$. Thus, the asymptotic relative efficiency of $\exp(\hat{\sigma}^2/2)$ with respect to $n^{-1}\sum_{i=1}^{n}\exp(e_i)$ is greater than one. This is expected, as a parametric estimate is generally more efficient than a nonparametric estimate when the parametric model is true.

### A.3 Bias correction factor for $\hat{C}$ with generalized Gaussian errors.

In Appendix A.1, a bias correction factor was derived for $\hat{C}$ under the assumption that the errors are normally distributed. In practice, the underlying distribution of the errors will not be exactly normal; it may be "more peaked" with "lighter tails", or "less peaked" with "heavier tails". For the data at hand, the distribution of the residuals appears to be approximately symmetric but with "heavier tails" than the normal. In this Appendix, the biasing factor $E(\exp(\epsilon))$ will be derived for the case of the errors being distributed according to the generalized Gaussian distribution

$$f(x) = \beta\ \exp(-|x/b|^p), \quad b > 0, \quad 1 < p \le 2, \tag{A3.1}$$

where $\beta$ is a normalizing constant. This is a family of symmetric distributions with heavier tails than the standard normal distribution.

Let $\epsilon$ be distributed according to (A3.1). By using gamma integrals, it can be shown that $\beta = p(2b\Gamma(1/p))^{-1}$. Because $f(x)$ is symmetric and has exponential tails, the odd moments of $\epsilon$ are zero. The even moments are

$$E(\epsilon^k) = 2\int_0^{\infty} \frac{px^k}{2b\Gamma(1/p)} \exp((-|x/k|^p)\, dx = b^k \frac{\Gamma((k+1)/p)}{\Gamma(1/p)}. \tag{A3.2}$$

So, assuming that the errors are generalized Gaussian, with $p$ known, the bias correction term is

$$E(\exp(\epsilon)) = \sum_{k=0}^{\infty} \frac{E(\epsilon^{2k})}{(2k!)} = \Big(\sum_{k=0}^{\infty} \frac{b^{2k}\Gamma((2k+1)/p)}{(2k)!}\Big)/\Gamma(1/p). \qquad (A2.3)$$

If $p > 1$, $E(\exp(\epsilon))$ is finite for all values of $b$ (this follows from the fact that if $p > 1$, $E(\exp(tX))$ exists for all t).

A study of the sensitivity of the bias correction term to the change in $p$ can be carried out as follow. Fix $\sigma^2 = \hat{\sigma}^2$, where $\hat{\sigma}^2$ is a residual mean square from one of the data sets. For each value of $p$ between 1 and 2, calculate $b$ that gives this $\sigma^2$, then substitute this $b$ in equation (A3.3) to obtain $E(\exp(\epsilon))$. Tabulate $p$ and $E(\exp(\epsilon))$ as a function of $p$ for several values of $p$ between 1 and 2. Such tables for Mission and Oldman are provided in Tables 4.6 and 4.7.

### A.4 Integration process for sediment load determination.

In this appendix, details of the integration process that is used to calculate the sediment load for the high-flow period are given.

As discussed in Section 3.3, linear interpolation between $\widehat{C_i}$'s and between $Q_i$'s is performed. Let

$$\widehat{C(t)} = \hat{a} + \hat{b}t$$

be the linear interpolation between $\widehat{C(t_1)}$ and $\widehat{C(t_2)}$, where $t_1 < t_2$, and

$$\widehat{Q(t)} = c + dt$$

be the linear interpolation between $Q(t_1)$ and $Q(t_2)$. To estimate the sediment load at time $t$ (see Section 1.2b), where $t_1 \le t \le t_2$, the following is used

$$\widehat{L(t)} = \widehat{C(t)}\widehat{Q(t)}$$
$$= (\hat{a} + \hat{b}t)(c + dt)$$

Thus the total sediment load between $t_1$ and $t_2$ is

$$\int_{t_1}^{t_2} \widehat{C(t)}\widehat{Q(t)}\,dt = \int_{t_1}^{t_2} (\hat{a} + \hat{b}t)(c + dt)\,dt$$
$$= \frac{\hat{b}d}{3}({t_2}^3 - {t_1}^3) + \frac{(\hat{a}d + \hat{b}d)}{2}({t_2}^2 - {t_1}^2) + \hat{a}c(t_2 - t_1).$$

Without loss of generality, let $t_1 = 0$. Then

$$\hat{a} = \widehat{C_1},\ c = Q_1,\ \hat{b} = (\widehat{C_2} - \widehat{C_1})/t_2 \text{ and } d = (Q_2 - Q_1)/t_2$$

and

$$\int_0^{t_2} C(t)Q(t)\,dt = \left[\frac{\widehat{C_2}Q_2}{2} - \frac{(\widehat{C_2} - \widehat{C_1})(Q_2 - Q_1)}{6} + \frac{\widehat{C_1}Q_1}{2}\right]t_2.$$

Recall from Section 1.2b that $Q$ is measured in m$^3$/s and $C$ is measured in mg/l. Hence, to estimate the sediment load for the first day, for example, $t_2$ is replaced by 86400, which is the number of seconds in a day.

The total sediment load estimate for the high-flow period is

$$\widehat{\text{Load}} = \sum_{i=2}^{n}\left[\frac{\widehat{C_i}Q_i}{2} - \frac{(\widehat{C_i} - \widehat{C_{i-1}})(Q_i - Q_{i-1})}{6} + \frac{\widehat{C_{i-1}}Q_{i-1}}{2}\right],$$

since $Q_i$ are obtained daily.

### A.5 Expectation and variance of the sediment load.

In this appendix, the expectation and the variance of sediment load for the high-flow period are derived assuming the model like (A5.1) below. These terms are needed for obtaining the standard error of the sediment load estimate.

As shown in Appendix A.4, the total sediment load for the high-flow period can be approximated by

$$\text{Load} = 86400 \sum_{i=2}^{n} \left[ \frac{C_i Q_i}{2} - \frac{(C_i - C_{i-1})(Q_i - Q_{i-1})}{6} + \frac{C_{i-1} Q_{i-1}}{2} \right].$$

This can be written as $\sum_{i=1}^{n} a_i C_i$, where $a_i = [(2Q_i/3) + (Q_{i-1}/6) + (Q_{i+1}/6)]$ for $i = 2 \ldots n-1$, $a_1 = (Q_2/6) + (Q_1/3)$ and $a_n = (Q_n/3) + (Q_{n-1}/6)$. To calculate the expectation and the variance of the above load, $E(C_i)$, $\text{Var}(C_i)$ and $\text{Cov}(C_i,\ C_{i+t})$ are needed (since $\text{Var}(\sum_{i=1}^{n} a_i C_i) = \sum_{i=1}^{n} a_i{}^2 \text{Var}(C_i) + 2 \sum \sum_{i \neq j} a_i a_j \text{Cov}(C_i,\ C_j)$).

Suppose for simplicity that

$$\ln C_i = \beta_0 + \beta_1 \ln Q_i + \epsilon_i, \qquad (A5.1)$$

where $\epsilon_i = \sim N(0,\ \sigma^2)$, $\epsilon_i = \rho \epsilon_{i-1} + \eta_i$, $\eta_i$ are independent and identically distributed random variables and $\delta^2 + \rho^2 \sigma^2 = \sigma^2$. Then

$$\ln C_i \sim N(\beta_0 + \beta_1 \ln Q_i,\ \sigma^2). \qquad (A5.2)$$

Hence, $E(C_i) = \exp(\beta_0 + \beta_1 \ln Q_i + \sigma^2/2)$, by evaluating the moment generating function (m.g.f.) of $\ln C_i$ at 1.

The variance of the high-flow period load is a function involving $\text{Var}(C_i) = E(C_i{}^2) - E(C_i)^2$ and $\text{Cov}(C_i, C_{i+t}) = E(C_i C_{i+t}) - E(C_i)E(C_{i+t})$. $E(C_i{}^2)$ can be obtained by evaluating the m.g.f. of $\ln C_i$ at 2., so that

$$\text{Var}(C_i) = \exp\Big( 2[\beta_0 + \beta_1 \ln Q_i + \sigma^2/2] \Big) - \exp\Big( 2[\beta_0 + \beta_1 \ln Q_i] + \sigma^2 \Big).$$

$E(C_i C_{i+t})$ can be calculated by noting that

$$\Big( \ln C_i, \ln C_{i+t} \Big) \sim N\left( \begin{pmatrix} E(\ln C_i) \\ E(\ln C_{i+t}) \end{pmatrix},\ \sigma^2 \begin{pmatrix} 1 & \rho^t \\ \rho^t & 1 \end{pmatrix} \right)$$

and that $E(C_iC_{i+t}) = E(\exp(\ln C_i)\ \exp(\ln C_{i+t}))$ is the joint moment generating function of the bivariate normal distribution with both arguments equal to 1. Hence, $E(C_iC_{i+t}) = \exp(\mu_i + \mu_{i+t} + (2\sigma^2(1+\rho^t))/2)$ and

$$\mathrm{Cov}(C_i, C_{i+t}) = \exp\Big(\mu_i + \mu_{i+t} + \sigma^2(1+\rho^t)\Big) - E(C_i)E(C_{i+t}),$$

where $\mu_i = E(\ln C_i)$, and $\mu_{i+t} = E(\ln C_{i+t})$ are given in (A4.2).

The variance of the load can be as indicated in the first paragraph of this appendix. The standard error of the load estimate can be obtained by replacing $E(\ln C_i)$, $\sigma^2$, and $\rho$, with their estimates in $\sqrt{\mathrm{Var(Load)}}$.

### A.6 Maximum likelihood estimators.

As mentioned in Section 3.2a, the unknown parameters in the models (3.5) and (3.6) can be estimated using the maximum likelihood method. However, as this method is computationally more difficult, an alternative method was used. Following is a discussion of the maximum likelihood estimates (m.l.e.) and how they can be obtained. These estimates are adaptations of those in Millard and Guttorp (1985).

Suppose for simplicity that there $n$ years of data with $T$ consecutive sampled days in each year and that the true model for $\ln C_t$ is

$$\ln C_{it} = \beta_0 + \beta_1 \ln Q_{it} + \epsilon_{it}, \quad \epsilon_{it} = \rho\epsilon_{i,t-1} + \eta_{it}, \quad t = 2, \ldots T, \quad i = 1, \ldots n, \qquad (A6.1)$$

where

$$\underline{\epsilon_i} = \begin{pmatrix} \epsilon_{i1} \\ \vdots \\ \epsilon_{iT} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}, \sigma^2 \begin{pmatrix} 1 & \rho & \ldots & \rho^{T-1} \\ \vdots & \vdots & \ddots & \vdots \\ \rho^{T-1} & \rho^{T-2} & \ldots & 1 \end{pmatrix}\right), \qquad (A6.2)$$

$\{\eta_t\}$ are iid $N(0,\ \delta^2)$ random variables and $\delta^2 + \rho^2\sigma^2 = \sigma^2$, and $\underline{\epsilon_i}$ are independent vectors.

The likelihood function is

$$L(\sigma^2,\ \rho,\ \beta_0,\ \beta_1) = \left((2\pi)^{nT/2}|V|^{1/2}\right)^{-1} \exp\left(-\frac{1}{2}\varepsilon' V^{-1}\varepsilon\right),$$

where

$$\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}, \quad \varepsilon_i = \begin{pmatrix} \varepsilon_{i1} \\ \vdots \\ \varepsilon_{iT} \end{pmatrix} = \begin{pmatrix} \ln C_{i1} - \beta_0 - \beta_1 \ln Q_{i1} \\ \vdots \\ \ln C_{iT} - \beta_0 - \beta_1 - \ln Q_{iT} \end{pmatrix}$$

and

$$V = \begin{pmatrix} \Sigma & 0 & \cdots & 0 \\ 0 & \Sigma & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \Sigma \end{pmatrix},$$

with $\Sigma$ being the covariance matrix in (A6.2). It should be noted that there are two cases to consider. In one case, only the data during the high-flow period are included in the analysis. In the other case, data over the whole year are analyzed. In the former case, $n$ and $T$ are defined as above. For the latter case, all the years on record can be joined into a sequence of observations, $T$ is the total number of days on record and $n$ is taken to be 1.

The maximum likelihood estimates of $\beta_0,\ \beta_1,\ \rho$ and $\sigma^2$, can be obtained by solving the derivative equations of the logarithm of the likelihood function with respect to these parameters. These equations are simplified to

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \left[X' A^{-1} X\right]^{-1} X' A^{-1} Y, \tag{A6.3}$$

$$\sigma^2 = [(1-\rho^2)nT]^{-1}\varepsilon' A^{-1}\varepsilon, \tag{A6.4}$$

and

$$\sum_{i=1}^{n}\sum_{j=2}^{T}\left(\varepsilon_{i,j}\varepsilon_{i,j-1} - \rho\varepsilon_{i,j-1}^2\right) = 0, \tag{A6.5}$$

where

$$X = \begin{pmatrix} 1 & \ln Q_{11} \\ \vdots & \vdots \\ 1 & \ln Q_{nT} \end{pmatrix}, \quad Y = \begin{pmatrix} \ln C_{11} \\ \vdots \\ \ln C_{nT} \end{pmatrix} \text{ and } A = \frac{V}{\sigma^2}.$$

Combining (A6.4) and (A6.5) and simplifying yield $\hat{\sigma}^2 \approx (nT)^{-1} \sum\sum \epsilon_{i,j}^2$. For these calculations, the identities $\det \Omega = (1-\rho^2)^n$ and

$$\Omega^{-1} = \begin{pmatrix} 1 & -\rho & 0 & \cdots & 0 & 0 \\ -\rho & (1+\rho^2) & -\rho & \cdots & 0 & 0 \\ 0 & -\rho & (1+\rho^2) & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & (1+\rho^2) & -\rho \\ 0 & 0 & 0 & \cdots & -\rho & 1 \end{pmatrix}$$

are used, where $\Omega = \Sigma/\sigma^2$. No explicit forms exist for the m.l.e.'s of these parameters, but they can be obtained using the following iterative algorithm.

1. Estimate $\beta_0$ and $\beta_1$ using ordinary least squares.

2. Estimate $\rho$ using the residuals obtained in the previous step with

$$\hat{\rho} = \sum_{i=1}^{n}\sum_{j=2}^{T} \varepsilon_{i,j}\varepsilon_{i,j-1} / \sum_{i=1}^{n}\sum_{j=2}^{T} \varepsilon_{i,j}^2.$$

3. Estimate $\beta_0$ and $\beta_1$ in (A6.3) by substituting $\rho$ with $\hat{\rho}$ in $A = A(\rho)$.

4. Estimate $\rho$ as in step 2 using the residuals obtained in step 3.

5. Repeat steps 3 and 4 until estimates of $\beta_0$, $\beta_1$ and $\rho$ converge.

6. Estimate $\sigma^2$ with $\hat{\sigma}^2 = (nT)^{-1} \sum_{i=1}^{n} \sum_{j=2}^{T} \varepsilon_{i,j}^2$.

When the normal model for the errors is true, the method of estimation described above is preferred over the one suggested in Chapter 3. When the error model is an approximation, the method used in Chapter 3 is preferred since it is probably more robust to the normality assumption. The relationship between sediment concentration and water discharge is not dependent on the error model.

$X_1$ The observed discharge on day $i$, $Q_i$.

$X_2$ The first-order estimate of the relative change of discharge on day $i$.

$X_3$ The time in days to the first hydrograph peak encountered by scanning backwards through the discharge record.

$X_4$ The difference in discharge between the largest flow of the year up to and including day $i-1$ and the flow on day $i$.

$X_5$ The number of days since the major spring rise, defined as the minimum $j$ for which $Q(j) > 5 \min_{1 \leq n < i} Q(n)$.

$X_6$ The number of days counted backward from day $i$ for which $Q(i-m) < Q(i)$, limited to a maximum of 731 days.

Table 2.1 The prediction factors investigated by Kellerhals *et al.*.

| submodel | $p$ | $C_p$ | adjusted $R^2$ |
|---|---|---|---|
| $\ln Q$ | 2 | 2010.2 | .66 |
| $\ln Q_{-10}$ | 2 | 4988.8 | .34 |
| $abs(rate)$ | 2 | 5892.8 | .24 |
| $dslp$ | 2 | 4561.8 | .38 |
| $prMadis$ | 2 | 4830.9 | .36 |
| $\ln Q,\ \ln Q_{-10}$ | 3 | 344.4 | .84 |
| $\ln Q,\ abs(rate)$ | 3 | 1774.5 | .69 |
| $\ln Q,\ dslp$ | 3 | 1841.6 | .68 |
| $\ln Q,\ prMadis$ | 3 | 505.7 | .82 |
| $\ln Q,\ \ln Q_{-10},\ abs(rate)$ | 4 | 328.1 | .84 |
| $\ln Q,\ \ln Q_{-10},\ dslp$ | 4 | 317.8 | .84 |
| $\ln Q,\ \ln Q_{-10},\ prMadis$ | 4 | 5.9 | .88 |
| $\ln Q,\ \ln Q_{-10},\ abs(rate),\ dslp$ | 5 | 303.0 | .85 |
| $\ln Q,\ \ln Q_{-10},\ abs(rate),\ prMadis$ | 5 | 4.0 | .88 |
| $\ln Q,\ \ln Q_{-10},\ dslp,\ prMadis$ | 5 | 7.5 | .88 |
| $\ln Q,\ \ln Q_{-10},\ abs(rate),\ dslp,\ prMadis$ | 6 | 5.6 | .88 |

Table 3.1 $C_p$ and adjusted $R^2$ statistics for some of the better rising stage submodels of Mission.

| submodel | $p$ | $C_p$ | adjusted $R^2$ |
|---|---|---|---|
| $\ln Q$ | 2 | 151.6 | .81 |
| $\ln Q_{-10}$ | 2 | 1379.8 | .66 |
| $abs(rate)$ | 2 | 5154.4 | .20 |
| $dslp$ | 2 | 1777.1 | .61 |
| $prMadis$ | 2 | 6676.4 | .02 |
| $\ln Q$, $\ln Q_{-10}$ | 3 | 56.2 | .82 |
| $\ln Q$, $abs(rate)$ | 3 | 118.1 | .81 |
| $\ln Q$, $dslp$ | 3 | 151.8 | .81 |
| $\ln Q$, $prMadis$ | 3 | 133.6 | .81 |
| $\ln Q$, $\ln Q_{-10}$, $abs(rate)$ | 4 | 29.2 | .82 |
| $\ln Q$, $\ln Q_{-10}$, $dslp$ | 4 | 39.1 | .82 |
| $\ln Q$, $\ln Q_{-10}$, $prMadis$ | 4 | 47.6 | .82 |
| $\ln Q$, $\ln Q_{-10}$, $abs(rate)$, $dslp$ | 5 | 9.3 | .83 |
| $\ln Q$, $\ln Q_{-10}$, $abs(rate)$, $prMadis$ | 5 | 17.9 | .83 |
| $\ln Q$, $\ln Q_{-10}$, $dslp$, $prMadis$ | 5 | 39.2 | .83 |
| $\ln Q$, $\ln Q_{-10}$, $abs(rate)$, $dslp$, $prMadis$ | 6 | 8.4 | .83 |

Table 3.2 $C_p$ and adjusted $R^2$ statistics for some of the better falling stage submodels of Mission.

| lag | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| $\widehat{\rho(i)}$ | .6819 | .5840 | .5017 | .4592 | .3808 | .3473 | .3276 | .2878 | .2785 | .2711 |
| $\widehat{pac(i)}$ | .6819 | .2224 | .0794 | .0843 | −.0258 | .0360 | .0527 | −.0090 | .0476 | .0390 |

Table 3.3 Sample autocorrelations and sample partial autocorrelations of the errors in the Mission model.

| order | $n\sum_{i=1}^{k}\ln(1-\widehat{pac(i)}^2)+2k$ |
|---|---|
| 1 | −1559 |
| 2 | −1684 |
| 3 | −1698 |
| 4 | −1713 |
| 5 | −1713 |
| 6 | −1713 |
| 7 | −1719 |
| 8 | −1717 |
| 9 | −1721 |
| 10 | −1723 |

Table 3.4 $AIC(k) - n\ln\hat{\sigma}^2$ used to select the order of the autoregressive model for the errors in the Mission model.

| station | rising stage model | falling stage model | error model |
|---|---|---|---|
| Mission | $\widehat{\ln C_i} = -0.40 + 3.40X_{1i} + 2.05X_{2i} - 1.40 \ln Q_{-10i}$ | $\widehat{\ln C_i} = -9.09 + 1.59 \ln Q_i$ | $\hat{\epsilon}_i = 0.83\epsilon_{i-1}$ |
| Hansard | $\widehat{\ln C_i} = -1.29 + 1.59 \ln Q_i - 0.66 \ln Q_{-10i}$ | $\widehat{\ln C_i} = -1.63 + 0.98 \ln Q_i$ | $\hat{\epsilon}_i = 0.76\epsilon_{i-1}$ |
| Brocket | $\widehat{\ln C_i} = 0.33 + 1.82 \ln Q_i - 1.00 \ln Q_{-8i}$ | $\widehat{\ln C_i} = -6.52 + 0.32Y_1 i + 2.09Y_2 i$ | $\hat{\epsilon}_i = 0.83\epsilon_{i-1}$ |
| Walsingham | $\widehat{\ln C_i} = 1.90 + 1.11 \ln Q_i$ | $\widehat{\ln C_i} = 3.45 + 2.03 \ln Q_i - 1.51 \ln Q_{-1i}$ | $\hat{\epsilon}_i = 0.70\epsilon_{i-1}$ |

*where* $Y1 = (\ln Q - 4) \times (1 - E)$ *and* $X1 = (\ln Q - 8) \times (1 - D)$

$Y2 = (\ln Q \times E) \times ((1 - D) \times 4)$ $X2 = (\ln Q \times D) \times ((1 - D) \times 8)$

$$E = \begin{cases} 0, \ if & \ln Q \le 4 \\ 1, \ if & \ln Q > 4 \end{cases} \qquad D = \begin{cases} 0, \ if & \ln Q \le 8 \\ 1, \ if & \ln Q > 8 \end{cases}$$

Table 3.5 Least squares fitted models.

| station | variable | least squares | min. abs. res. | Huber | Andrews |
|---|---|---|---|---|---|
| Mission | intercept | -0.40 | -0.68 | -0.51 | -0.52 |
| | X1 | 3.40 | 3.42 | 3.41 | 3.43 |
| | X2 | 2.05 | 2.09 | 2.05 | 2.05 |
| | $\ln Q_{-10}$ | -1.40 | -1.41 | -1.38 | -1.38 |
| Hansard | intercept | -1.29 | -1.00 | -1.30 | -1.31 |
| | $\ln Q$ | 1.59 | 1.55 | 1.58 | 1.58 |
| | $\ln Q_{-10}$ | -0.66 | -0.66 | -0.64 | -0.64 |
| Brocket | intercept | 0.33 | 0.51 | 0.34 | 0.36 |
| | $\ln Q$ | 1.82 | 1.97 | 1.89 | 1.91 |
| | $\ln Q_{-8}$ | -1.00 | -1.23 | -1.08 | -1.11 |
| Walsingham | intercept | 1.90 | 1.91 | 1.91 | 1.92 |
| | $\ln Q$ | 1.11 | 1.09 | 1.10 | 1.09 |

Table 3.6 Rising stage models' coefficients estimated using least squares, minimum absolute residual, Huber and Andrews regressions.

| station | variable | least squares | min. abs. res. | Huber | Andrews |
|---|---|---|---|---|---|
| Mission | intercept | -9.09 | -8.97 | -8.93 | -8.93 |
| | $\ln Q$ | 1.59 | 1.58 | 1.57 | 1.57 |
| Hansard | intercept | -1.63 | -1.07 | -1.42 | -1.35 |
| | $\ln Q$ | 0.98 | 0.89 | 0.95 | 0.94 |
| Brocket | intercept | -6.52 | -6.73 | -6.62 | -6.65 |
| | Y1 | 0.32 | 0.26 | 0.28 | 0.29 |
| | Y2 | 2.09 | 2.13 | 2.11 | 2.12 |
| Walsingham | intercept | 3.45 | 3.46 | 3.45 | 3.45 |
| | $\ln Q$ | 2.03 | 1.83 | 1.92 | 1.91 |
| | $\ln Q_{-1}$ | -1.51 | -1.32 | -1.41 | -1.40 |

Table 3.7 Falling stage models' coefficients estimated using least squares, minimum absolute residual, Huber and Andrews regressions.

| station | rising stage model | falling stage model | error model |
|---|---|---|---|
| Mission | $\ln C_i = \begin{cases} \beta_0 + \beta_1 \ln Q_i + \beta_2 \ln Q_{-10i} + \epsilon_i, & \text{for } \ln Q_i > 8 \\ \beta_0' + \beta_1' \ln Q_i + \beta_2 \ln Q_{-10i} + \epsilon_i, & \text{for } \ln Q_i \leq 8 \end{cases}$<br>*subject to continuity at* $\ln Q_i = 8$ | $\ln C_i = \beta_0 + \beta_1 \ln Q_i + \epsilon_i$ | $\epsilon_i = \rho\epsilon_{i-1} + \eta_i$ |
| Hansard | $\ln C_i = \beta_0 + \beta_1 \ln Q_i + \beta_2 \ln Q_{-10i} + \epsilon_i$ | $\ln C_i = \beta_0 + \beta_1 \ln Q_i + \epsilon_i$ | $\epsilon_i = \rho\epsilon_{i-1} + \eta_i$ |
| Brocket | $\ln C_i = \beta_0 + \beta_1 \ln Q_i + \beta_2 \ln Q_{-8i} + \epsilon_i$ | $\ln C_i = \begin{cases} \beta_0 + \beta_1 \ln Q_i + \epsilon_i, & \text{for } \ln Q_i > 4 \\ \beta_0' + \beta_1' \ln Q_i + \epsilon_i, & \text{for } \ln Q_i \leq 4 \end{cases}$<br>*subject to continuity at* $\ln Q_i = 4$ | $\epsilon_i = \rho\epsilon_{i-1} + \eta_i$ |
| Walsingham | $\ln C_i = \beta_0 + \beta_1 \ln Q_i + \epsilon_i$ | $\ln C_i = \beta_0 + \beta_1 \ln Q_i + \beta_2 \ln Q_{-1i} + \epsilon_i$ | $\epsilon_i = \rho\epsilon_{i-1} + \eta_i$ |

Table 3.8 The proposed models for Mission, Hansard, Brocket and Walsingham.

| station | $\left(\sum \exp(e_i)\right)/n$ | $\exp(\hat{\sigma}^2/2)$ |
|---|---|---|
| Mission | 1.0661 | 1.0626 |
| Hansard | 1.1123 | 1.1067 |
| Brocket | 1.4618 | 1.3893 |
| Walsingham | 1.2712 | 1.2458 |

Table 3.9 Numerical comparison between two estimates of $\exp(\sigma^2/2)$.

| year | estimated annual load | published annual load | standard error | c.v. |
|---|---|---|---|---|
| 1966 | $21.5252 \times 10^6$ | $22.7491 \times 10^6$ | $2.3201 \times 10^6$ | 10.77 |
| 1969 | $15.8766 \times 10^6$ | $17.3299 \times 10^6$ | $1.7522 \times 10^6$ | 11.03 |
| 1972 | $31.5177 \times 10^6$ | $34.1606 \times 10^6$ | $3.6349 \times 10^6$ | 11.53 |
| 1974 | $25.4903 \times 10^6$ | $27.5183 \times 10^6$ | $2.8044 \times 10^6$ | 11.0 |
| 1976 | $26.8747 \times 10^6$ | $27.4562 \times 10^6$ | $2.6212 \times 10^6$ | 9.75 |
| 1979 | $12.3572 \times 10^6$ | $15.0096 \times 10^6$ | $1.4485 \times 10^6$ | 11.72 |
| 1982 | $21.2988 \times 10^6$ | $25.5634 \times 10^6$ | $2.3096 \times 10^6$ | 10.84 |

Table 4.1 Estimated loads vs. published data for Fraser at Mission.
(c.v. is the coefficient of variation).

| year | estimated annual load | published annual load | standard error | c.v. |
|---|---|---|---|---|
| 1972 | $4.64268 \times 10^6$ | $4.787251 \times 10^6$ | $0.69344 \times 10^6$ | 14.90 |
| 1974 | $3.84450 \times 10^6$ | $2.90580 \times 10^6$ | $0.53203 \times 10^6$ | 13.83 |
| 1976 | $3.64673 \times 10^6$ | $3.36528 \times 10^6$ | $0.40598 \times 10^6$ | 11.13 |
| 1978 | $1.57306 \times 10^6$ | $1.75841 \times 10^6$ | $0.17586 \times 10^6$ | 11.17 |
| 1980 | $1.98128 \times 10^6$ | $1.72463 \times 10^6$ | $0.21311 \times 10^6$ | 10.75 |
| 1981 | $2.21504 \times 10^6$ | $1.58250 \times 10^6$ | $0.31622 \times 10^6$ | 14.27 |
| 1982* | $4.10280 \times 10^6$ | $4.26429 \times 10^6$ | $0.51429 \times 10^6$ | 12.53 |

Table 4.2 Estimated loads vs. published data for Fraser at Hansard.
(c.v. is the coefficient of variation).

* figures given for period between January and September only.

| year | estimated annual load | published annual load | standard error | c.v. |
|---|---|---|---|---|
| 1967 | 825659 | 996061 | 392730 | 47.56 |
| 1968 | 82606 | 85679 | 34615 | 41.90 |
| 1969 | 248710 | 274020 | 110001 | 44.22 |
| 1972 | 596555 | 753428 | 270590 | 45.35 |
| 1974 | 291576 | 306314 | 108947 | 37.36 |
| 1975 | 785850 | 1222067 | 516770 | 65.75 |
| 1978 | 143026 | 221912 | 47477 | 33.19 |

Table 4.3 Estimated loads vs. published data for Oldman near Brocket.
(c.v. is the coefficient of variation).

| year | estimated annual load | published annual load | standard error | c.v. |
|---|---|---|---|---|
| 1968 | 26585 | 31258 | 6400 | 24.07 |
| 1971 | 13747 | 14813 | 2433 | 17.69 |
| 1973 | 28807 | 24747 | 5306 | 18.41 |
| 1977 | 31628 | 37281 | 5111 | 16.15 |
| 1979 | 36017 | 30422 | 6318 | 17.54 |
| 1981 | 22241 | 20476 | 3570 | 16.05 |
| 1983 | 26942 | 23323 | 3076 | 11.41 |

Table 4.4 Estimated loads vs. published data for Big Creek near Walsingham.
(c.v. is the coefficient of variation).

| station | year | cross-validation estimate | estimate based on all years |
|---|---|---|---|
| Mission | 1966 | $21.2879 \times 10^6$ | $21.0585 \times 10^6$ |
|  | 1972 | $31.1083 \times 10^6$ | $31.1629 \times 10^6$ |
|  | 1979 | $11.8456 \times 10^6$ | $12.0256 \times 10^6$ |
| Hansard | 1974 | $3.9557 \times 10^6$ | $3.8235 \times 10^6$ |
|  | 1978 | $1.5295 \times 10^6$ | $1.5451 \times 10^6$ |
|  | 1982 | $4.1303 \times 10^6$ | $4.1641 \times 10^6$ |
| Brocket | 1967 | 822611 | 824675 |
|  | 1972 | 581562 | 595551 |
|  | 1978 | 139010 | 142139 |
| Walsingham* | 1973 | 29078 | 28807 |
|  | 1979 | 36782 | 36017 |
|  | 1983 | 27306 | 26942 |

Table 4.5 Cross-validation estimates vs. estimates based on all years.

* figures given for the whole year.

| $p$ | 1.1 | 1.2 | 1.3 | 1.4 | 1.5 | 1.6 | 1.7 | 1.8 | 1.9 | 2.0 |
|---|---|---|---|---|---|---|---|---|---|---|
| $b$ | 0.28590 | 0.32132 | 0.35291 | 0.38095 | 0.40577 | 0.42771 | 0.44713 | 0.46433 | 0.47958 | 0.49313 |
| $E(\widehat{\exp(\epsilon)})$ | 1.06422 | 1.06384 | 1.06357 | 1.06335 | 1.06318 | 1.06304 | 1.06293 | 1.06283 | 1.06275 | 1.06268 |

Table 4.6 Sensitivity of the estimated bias correction factor to the change in the generalized Gaussian parameter $p$ for Mission.

| $p$ | 1.1 | 1.2 | 1.3 | 1.4 | 1.5 | 1.6 | 1.7 | 1.8 | 1.9 | 2.0 |
|---|---|---|---|---|---|---|---|---|---|---|
| $b$ | 0.66489 | 0.74726 | 0.82073 | 0.88593 | 0.94365 | 0.99469 | 1.03984 | 1.07983 | 1.11530 | 1.14682 |
| $E(\widehat{\exp(\epsilon)})$ | 1.45798 | 1.43859 | 1.42546 | 1.41597 | 1.40879 | 1.40318 | 1.39867 | 1.39498 | 1.39190 | 1.38930 |

Table 4.7 Sensitivity of the estimated bias correction factor to the change in generalized Gaussian parameter $p$ for Brocket.

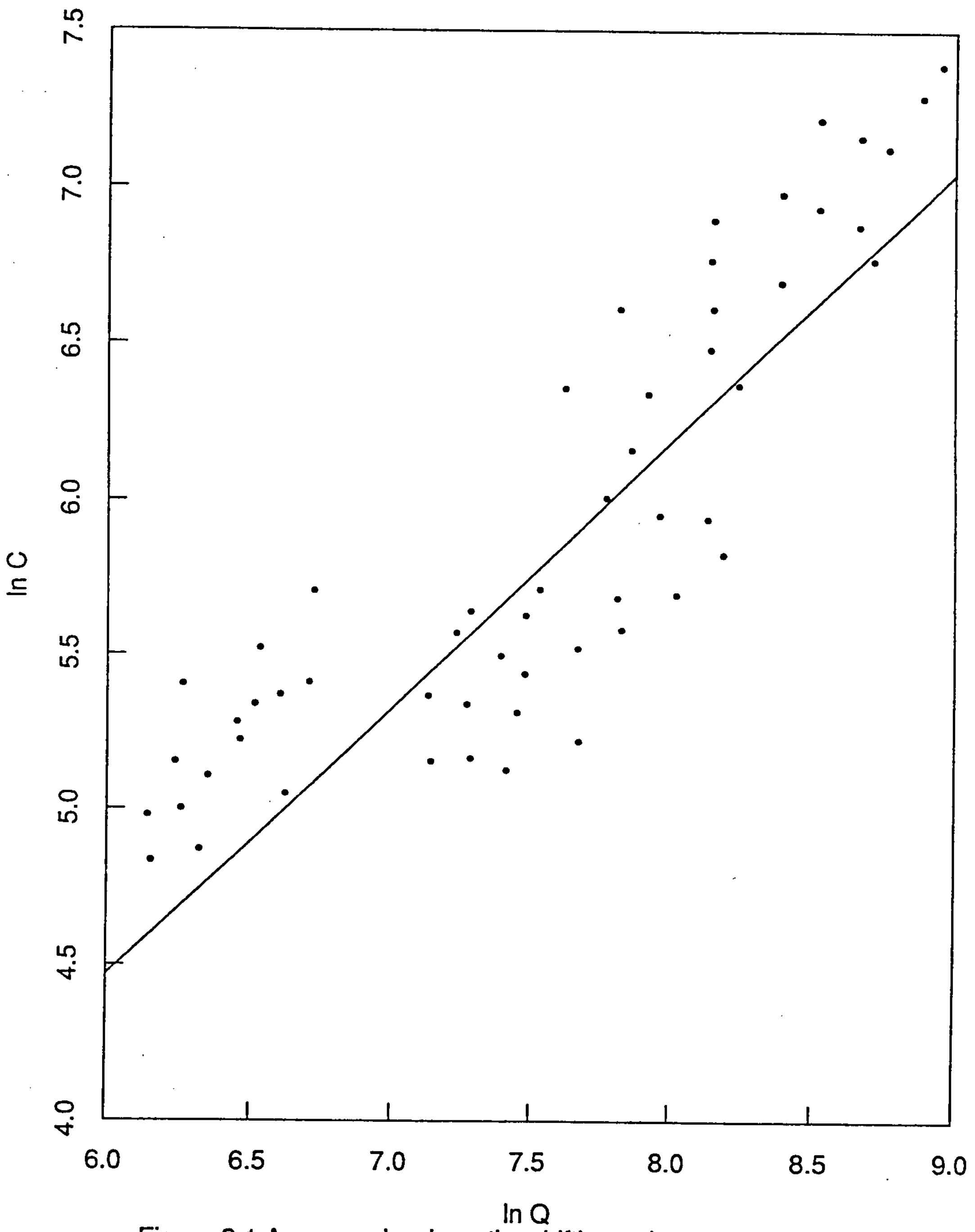

Figure 2.1 An example where the shifting rating curve can be used.

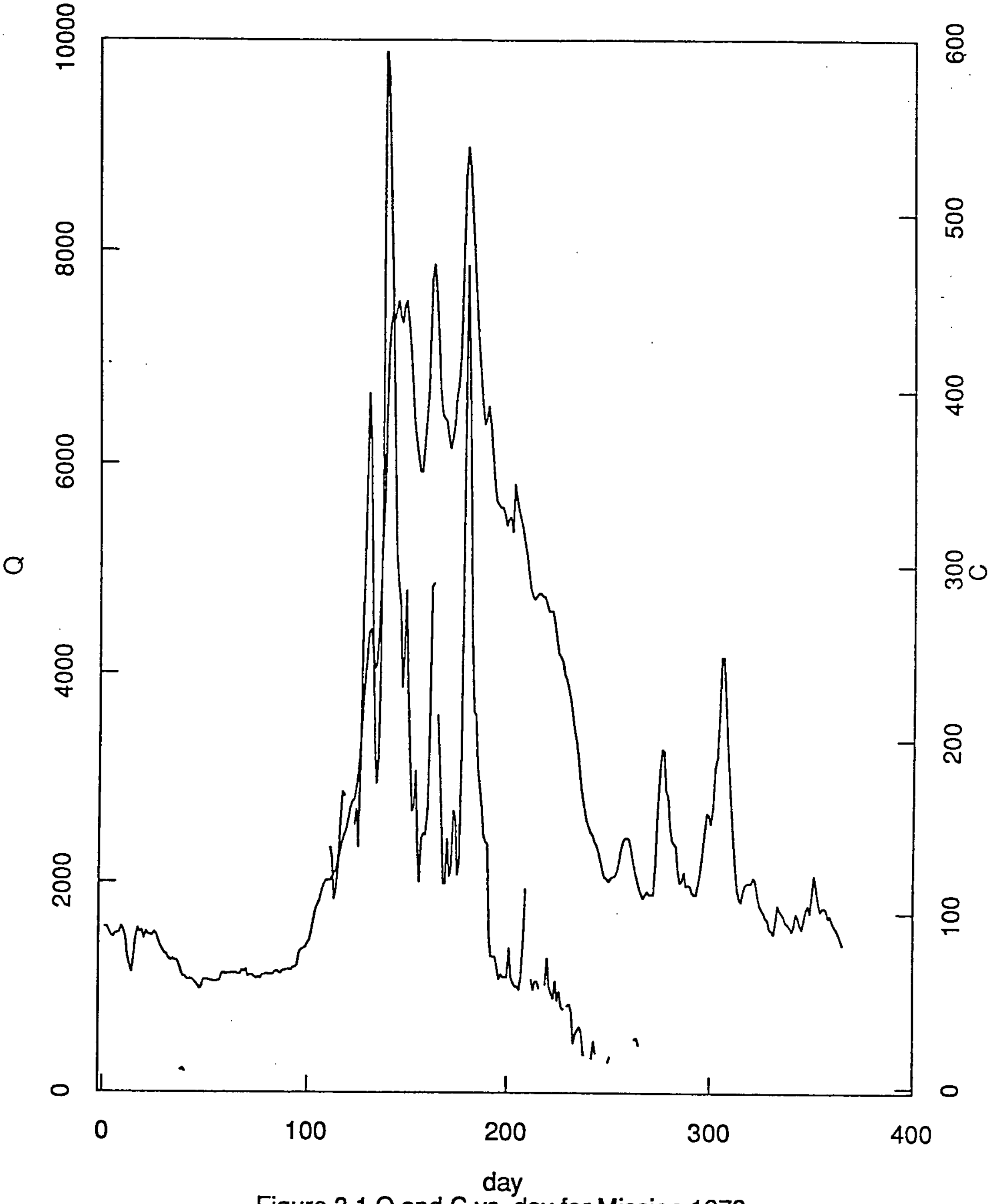

Figure 3.1 Q and C vs. day for Mission 1973.

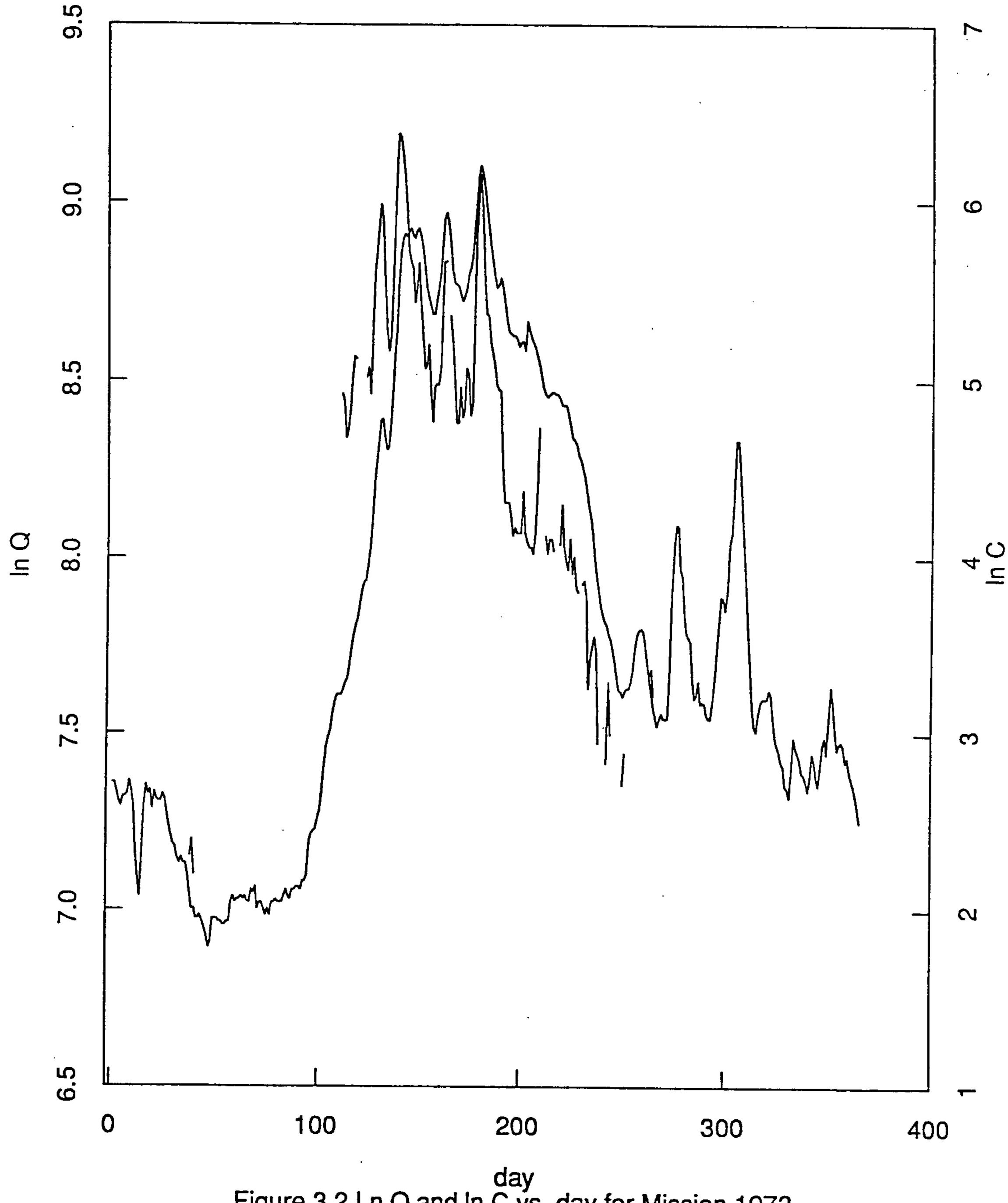

Figure 3.2 Ln Q and ln C vs. day for Mission 1973.

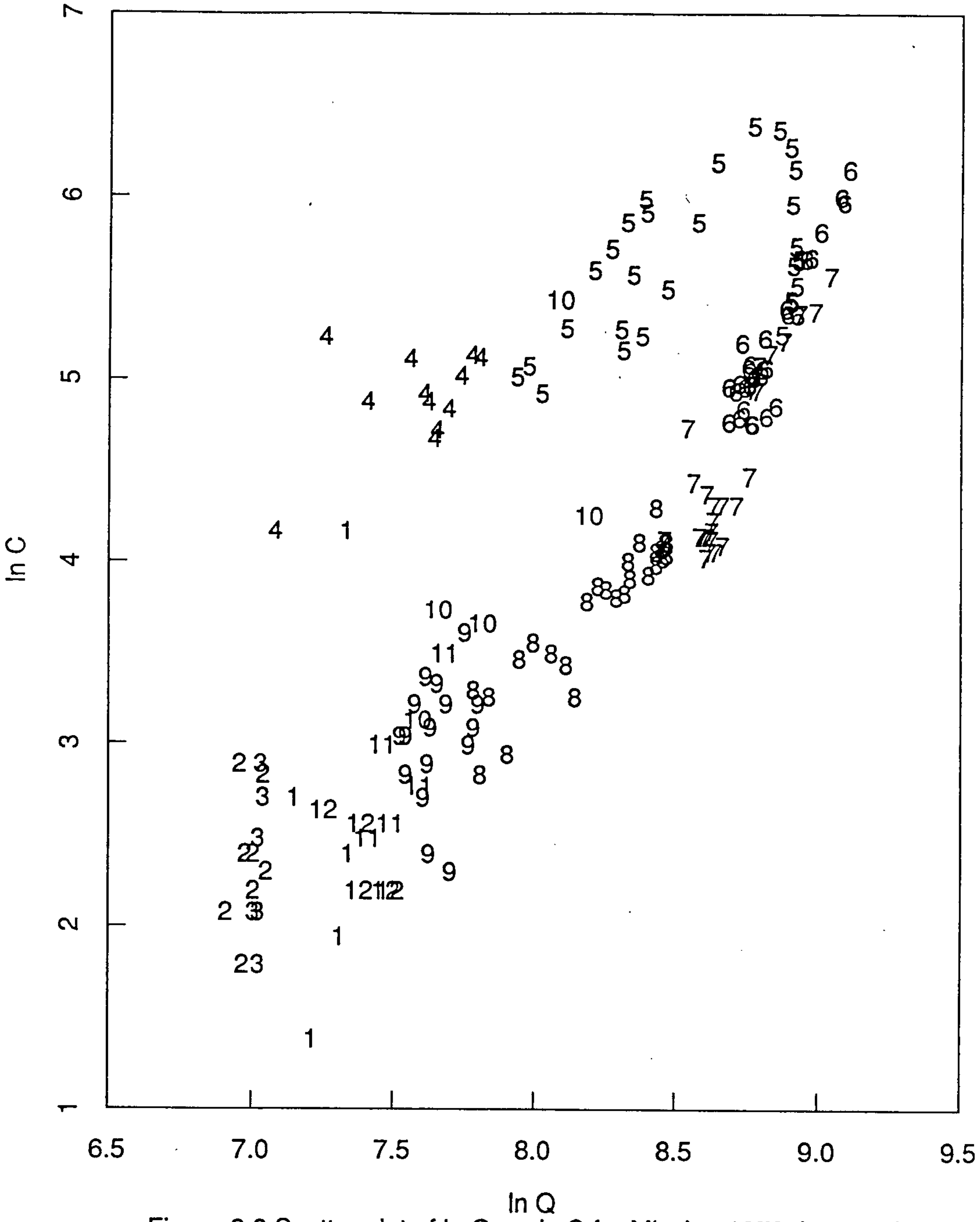

Figure 3.3 Scatter plot of ln Q vs. ln C for Mission 1973, by month.

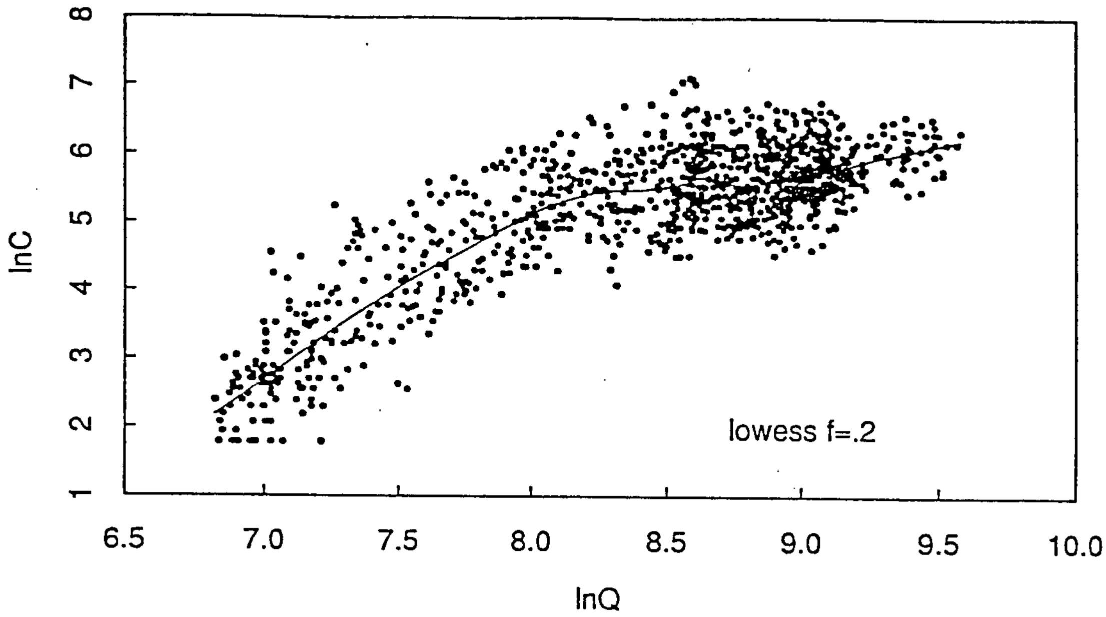

Figure 3.4a Scatter plot of $\ln Q$ vs. $\ln C$ for the rising stage of Mission.

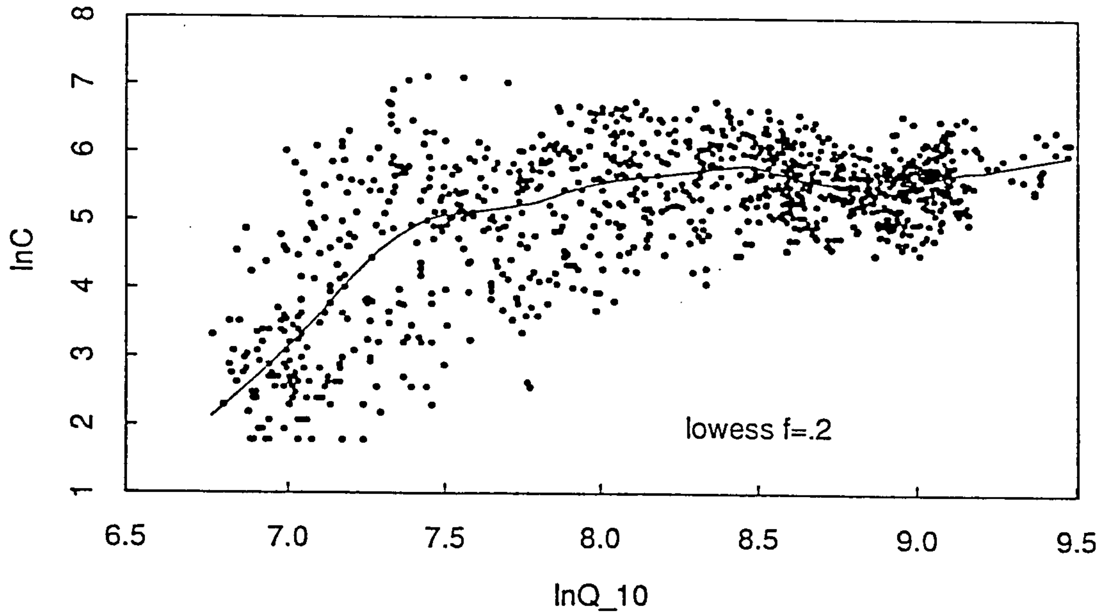

Figure 3.4b Scatter plot of $\ln Q_{-10}$ vs. $\ln C$ for the rising stage of Mission.

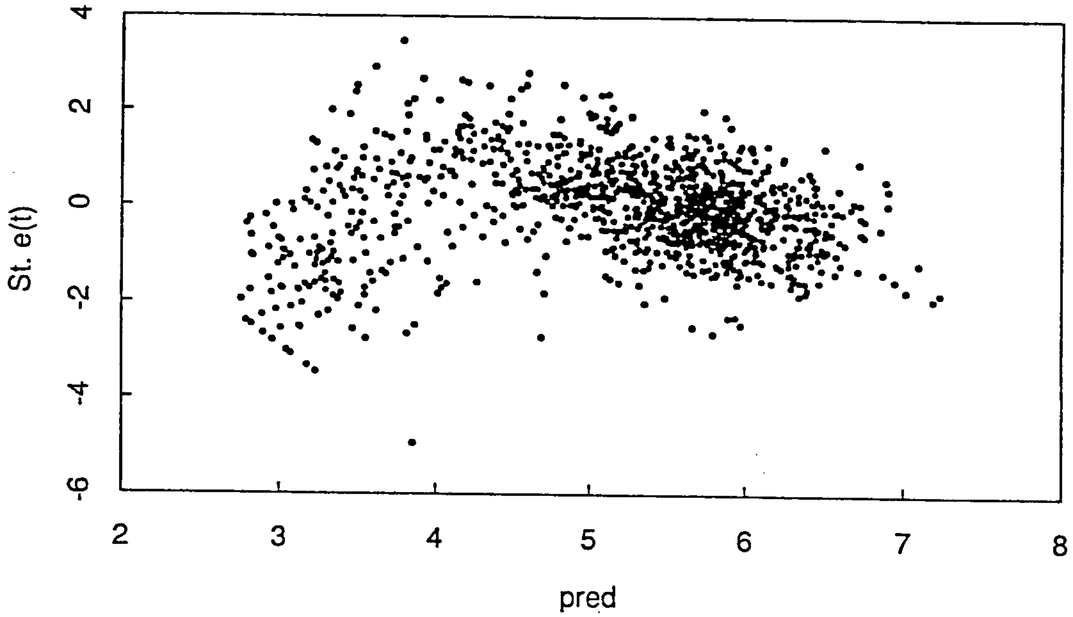

Figure 3.5a Predicted $\ln C$ vs. Studentized residual for the rising stage of Mission.

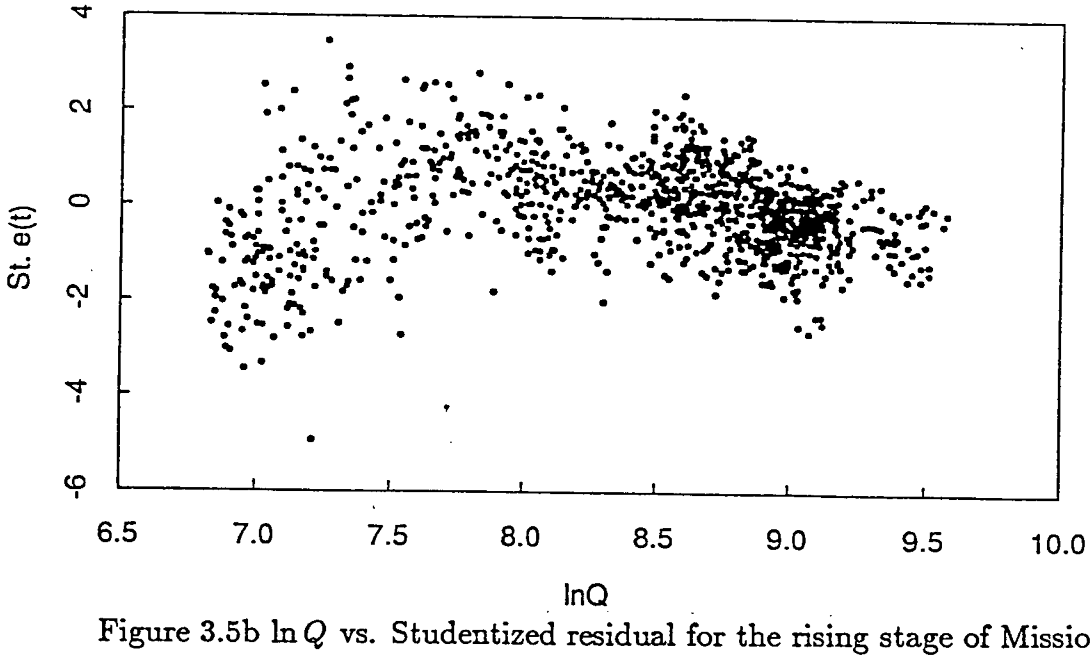

Figure 3.5b $\ln Q$ vs. Studentized residual for the rising stage of Mission.

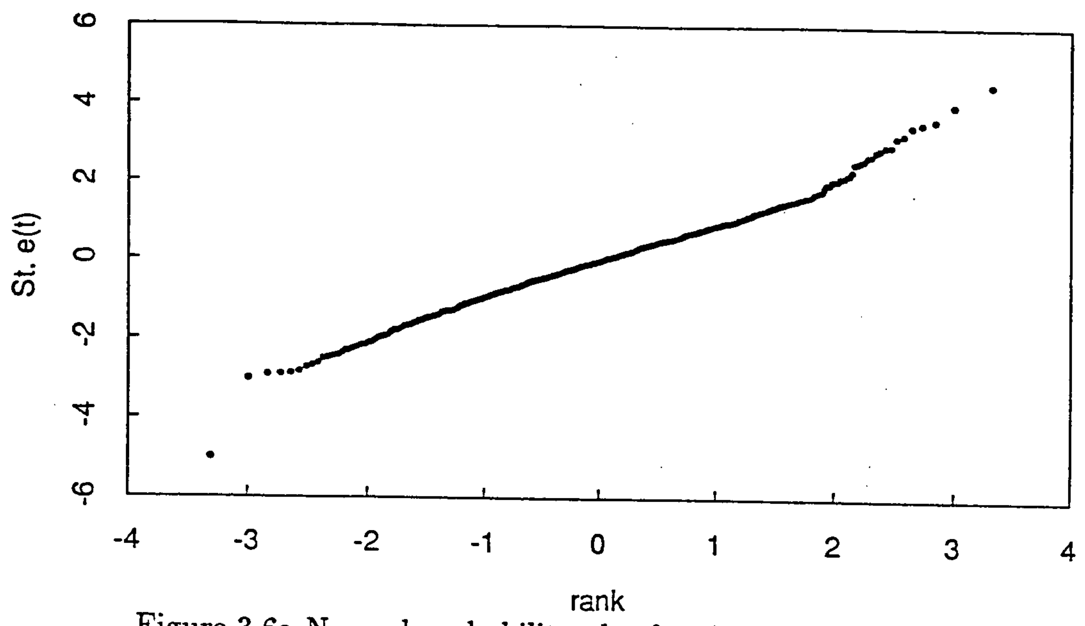

Figure 3.6a Normal probability plot for the residuals of the rising stage, Mission.

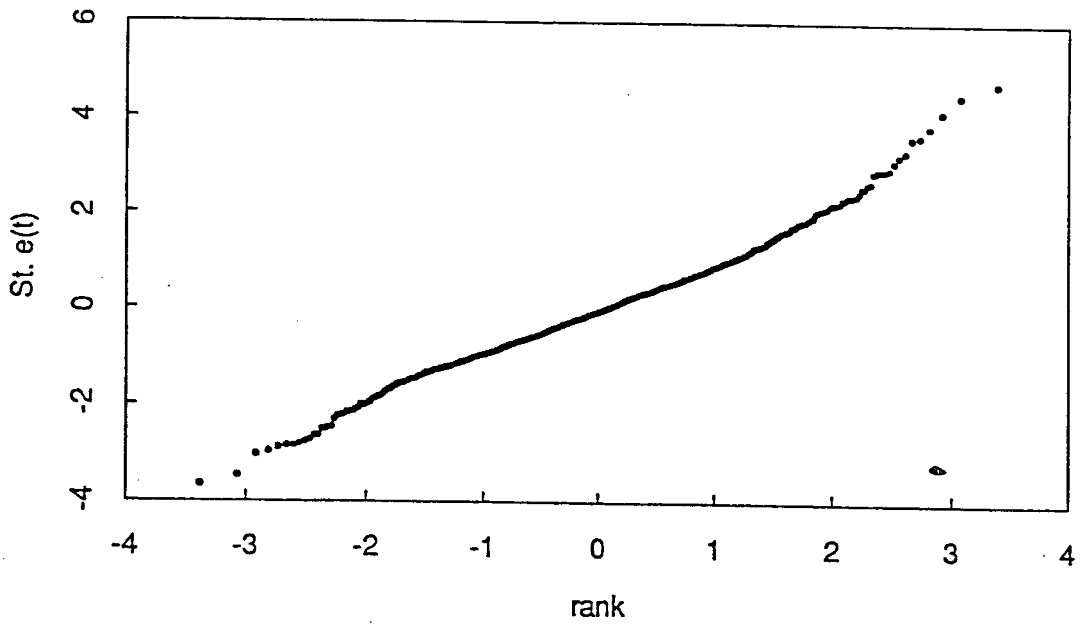

Figure 3.6b Normal probability plot for the residuals of the falling stage, Mission.

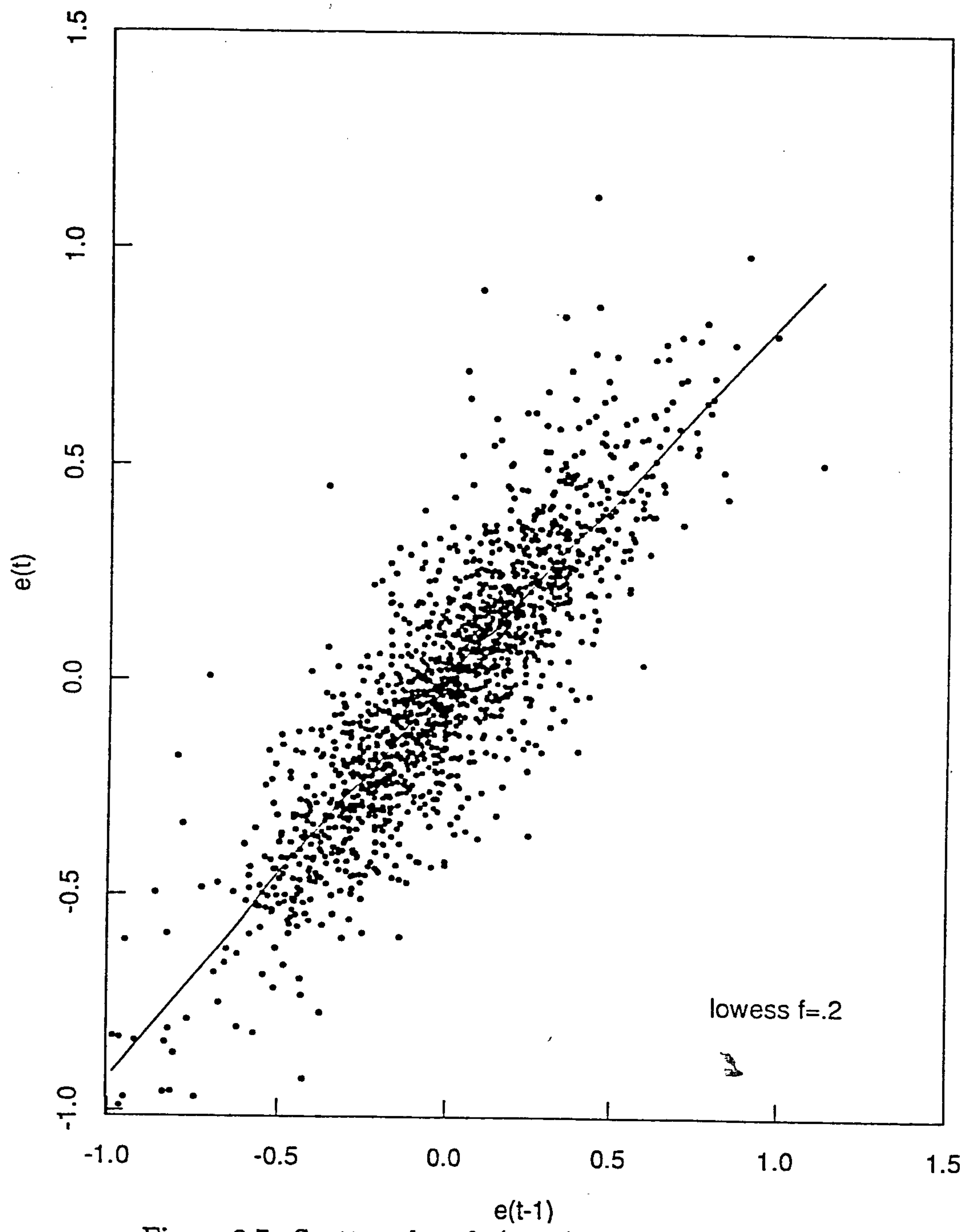

Figure 3.7a Scatter plot of $e(t-1)$ vs. $e(t)$ for Mission.

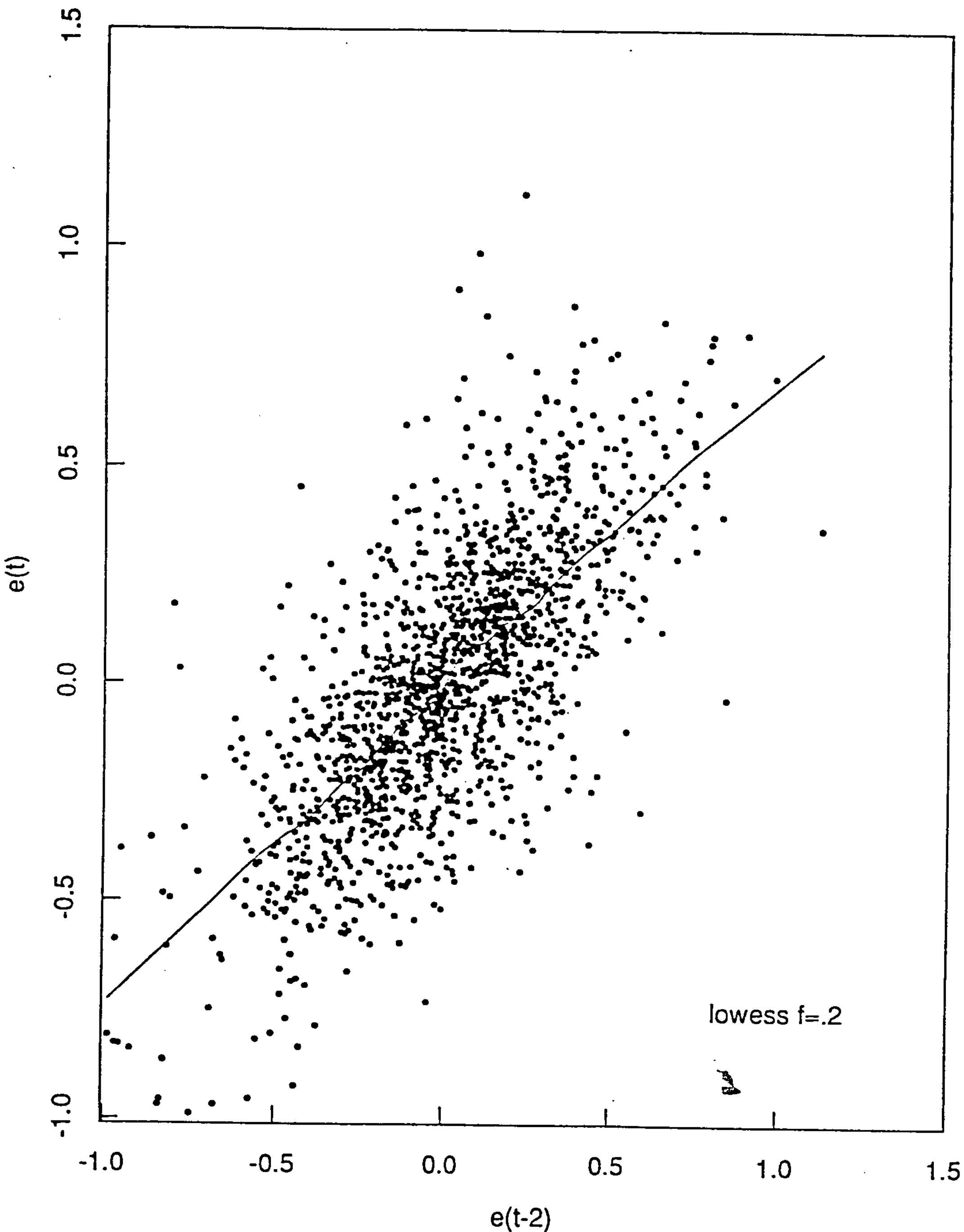

Figure 3.7b Scatter plot of $e(t-2)$ vs. $e(t)$ for Mission.

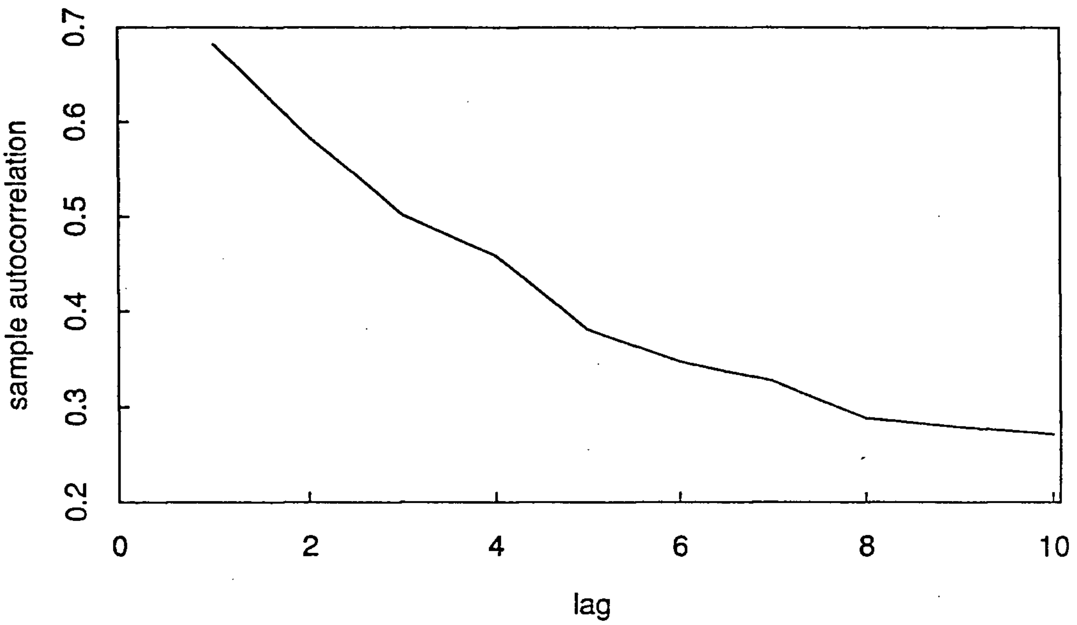

Figure 3.8 Lag vs. sample autocorrelation of residuals, Mission.

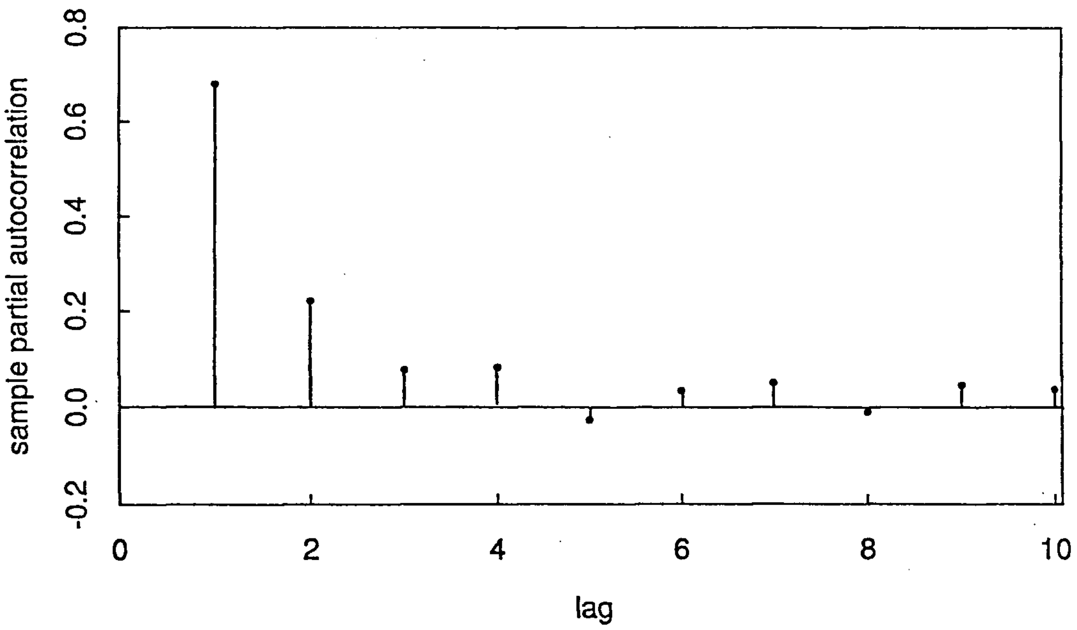

Figure 3.9 Lag vs. sample partial autocorrelation of residuals, Mission.

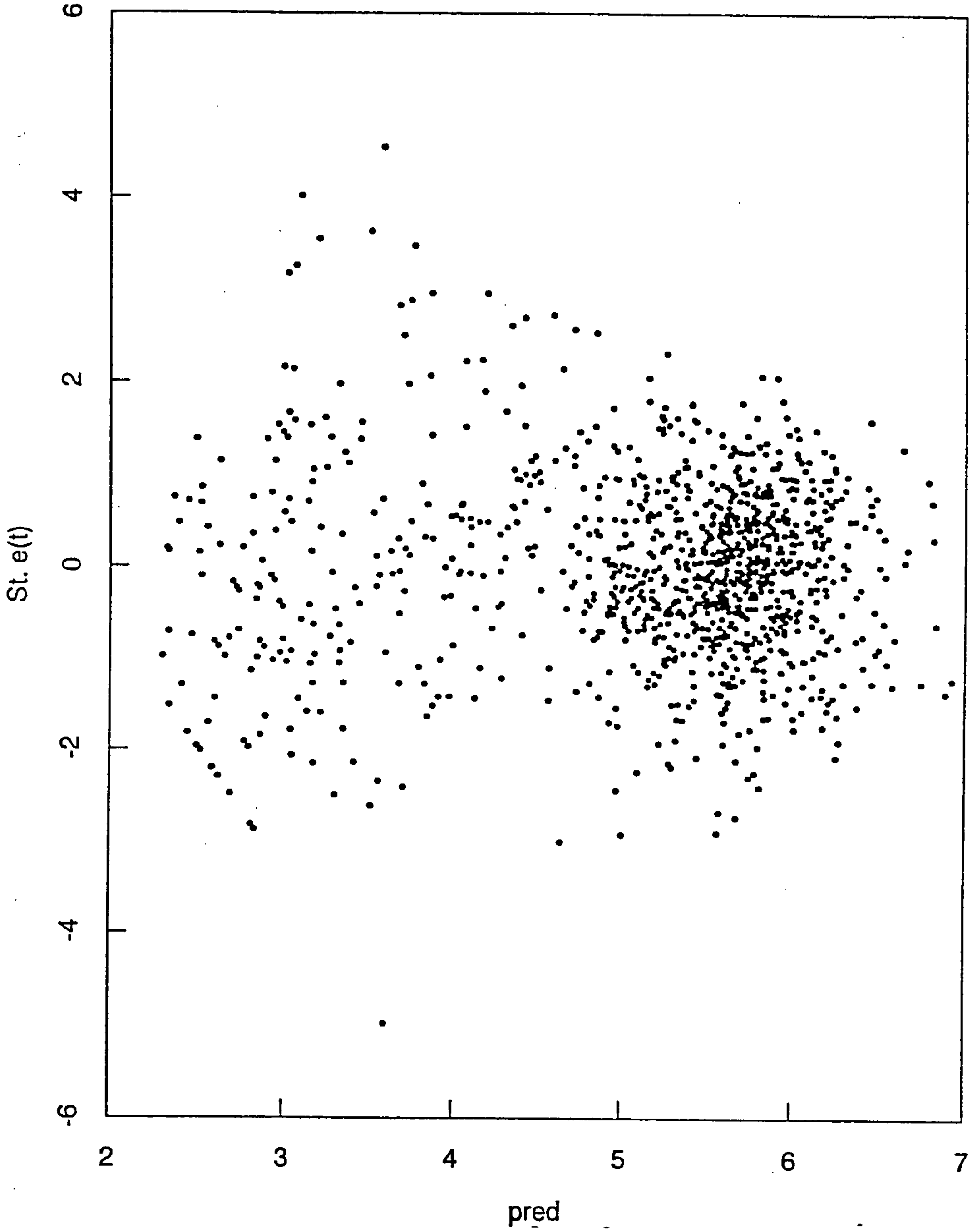

Figure 3.10a Predicted $\ln C$ vs. Studentized residual for the rising stage of Mission.

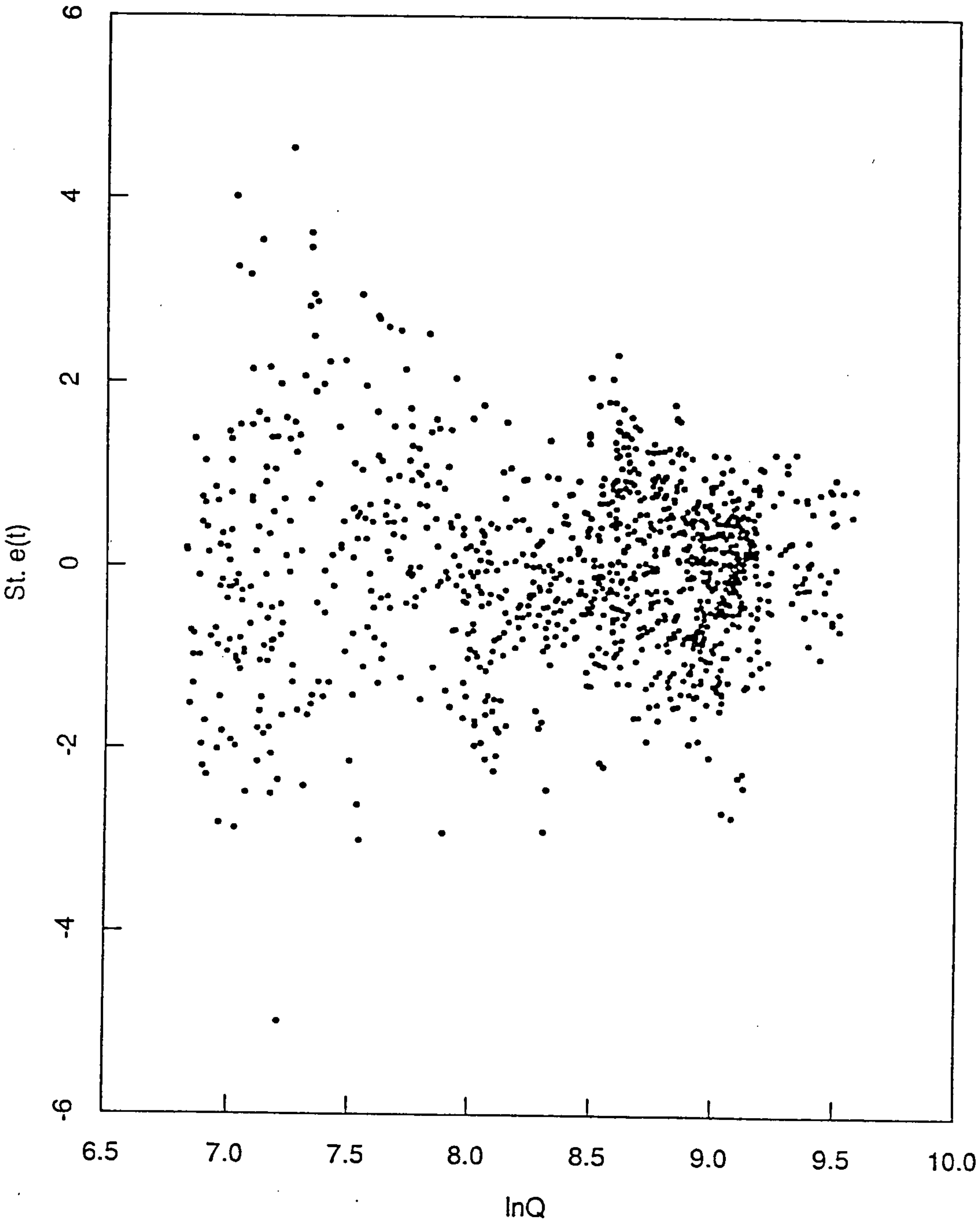

Figure 3.10b $\ln Q$ vs. Studentized residual for the rising stage of Mission.

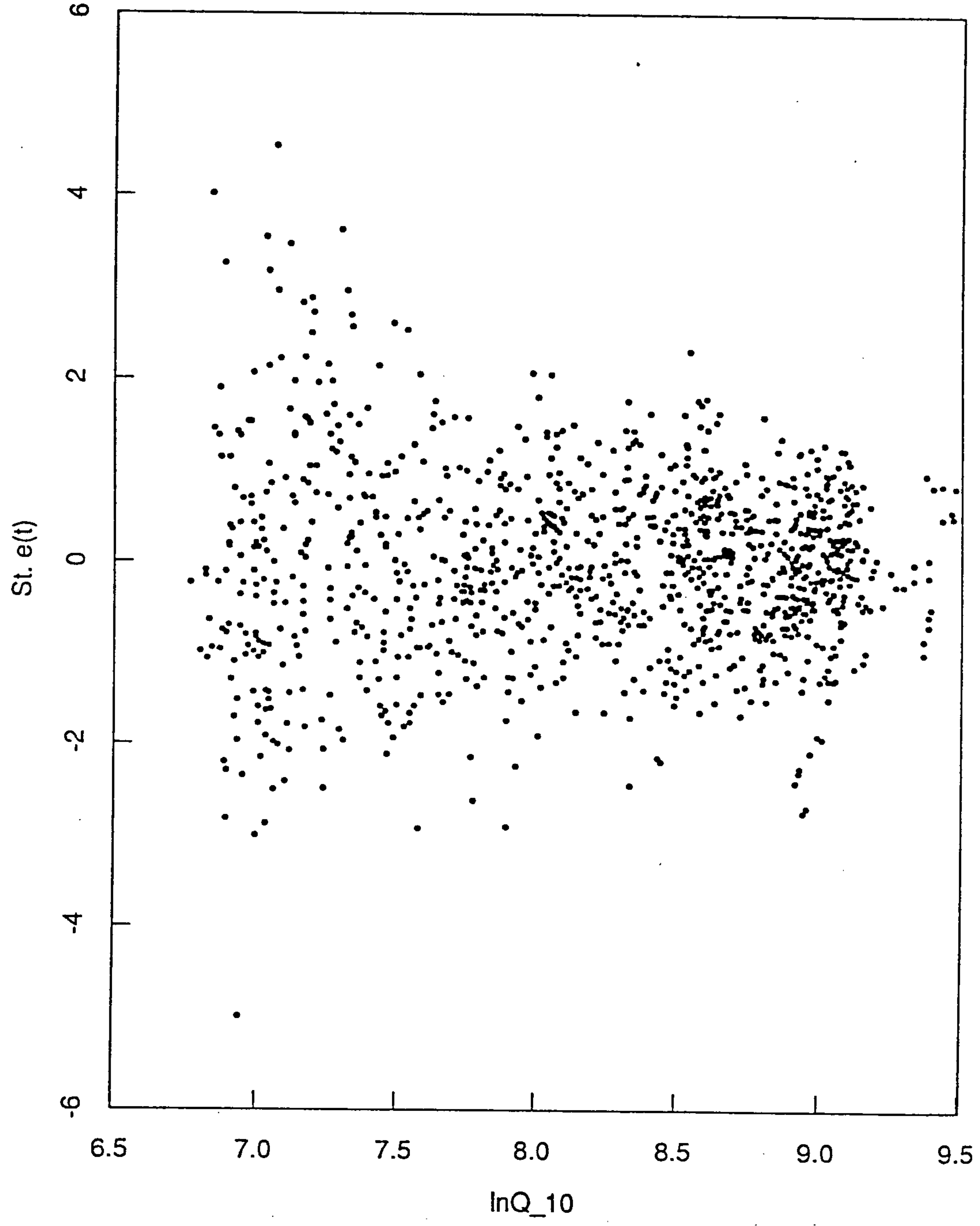

Figure 3.10c $\ln Q_{-10}$ vs. Studentized residual for the rising stage of Mission.

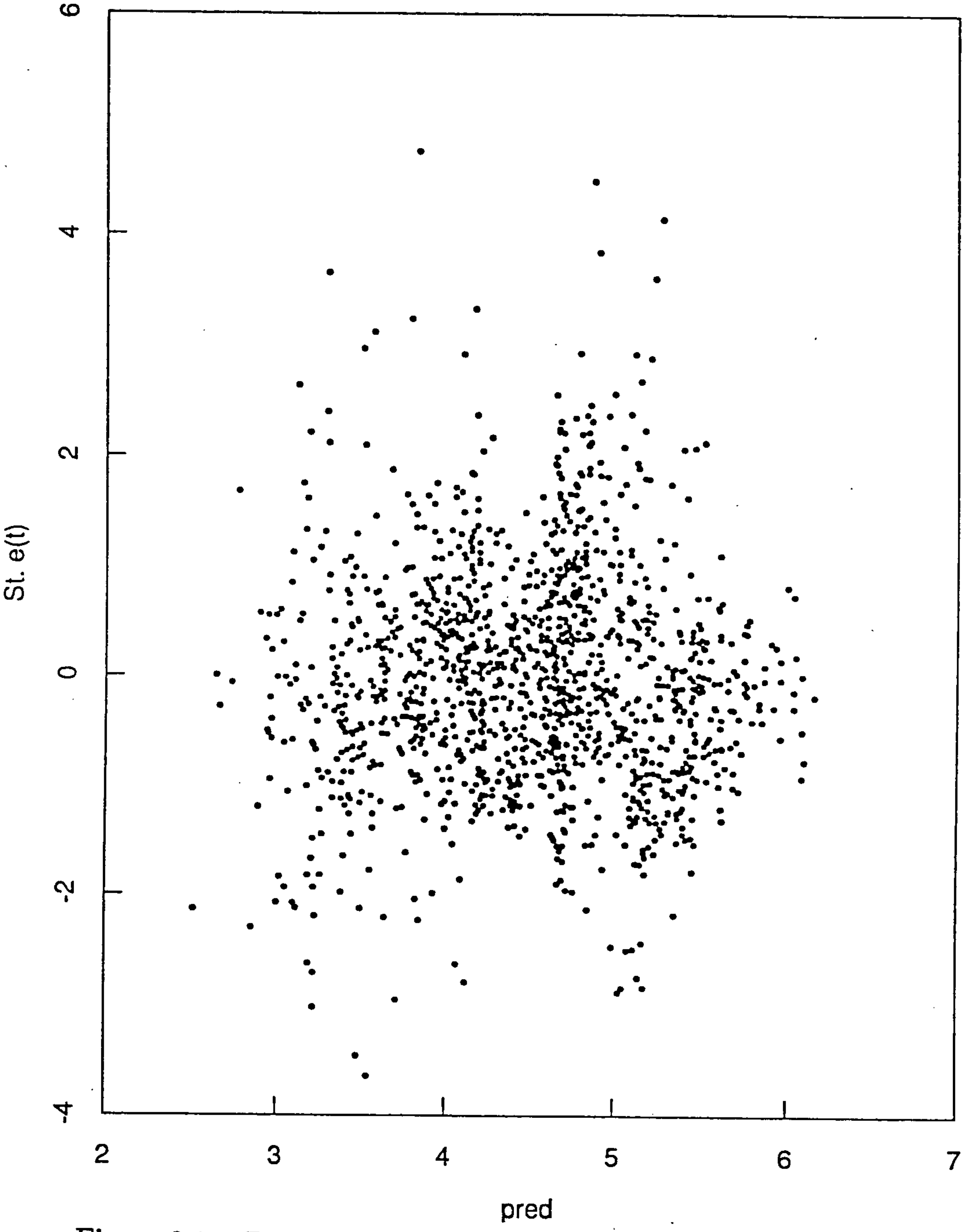

Figure 3.11a Predicted $\ln C$ vs. Studentized residual for the falling stage of Mission.

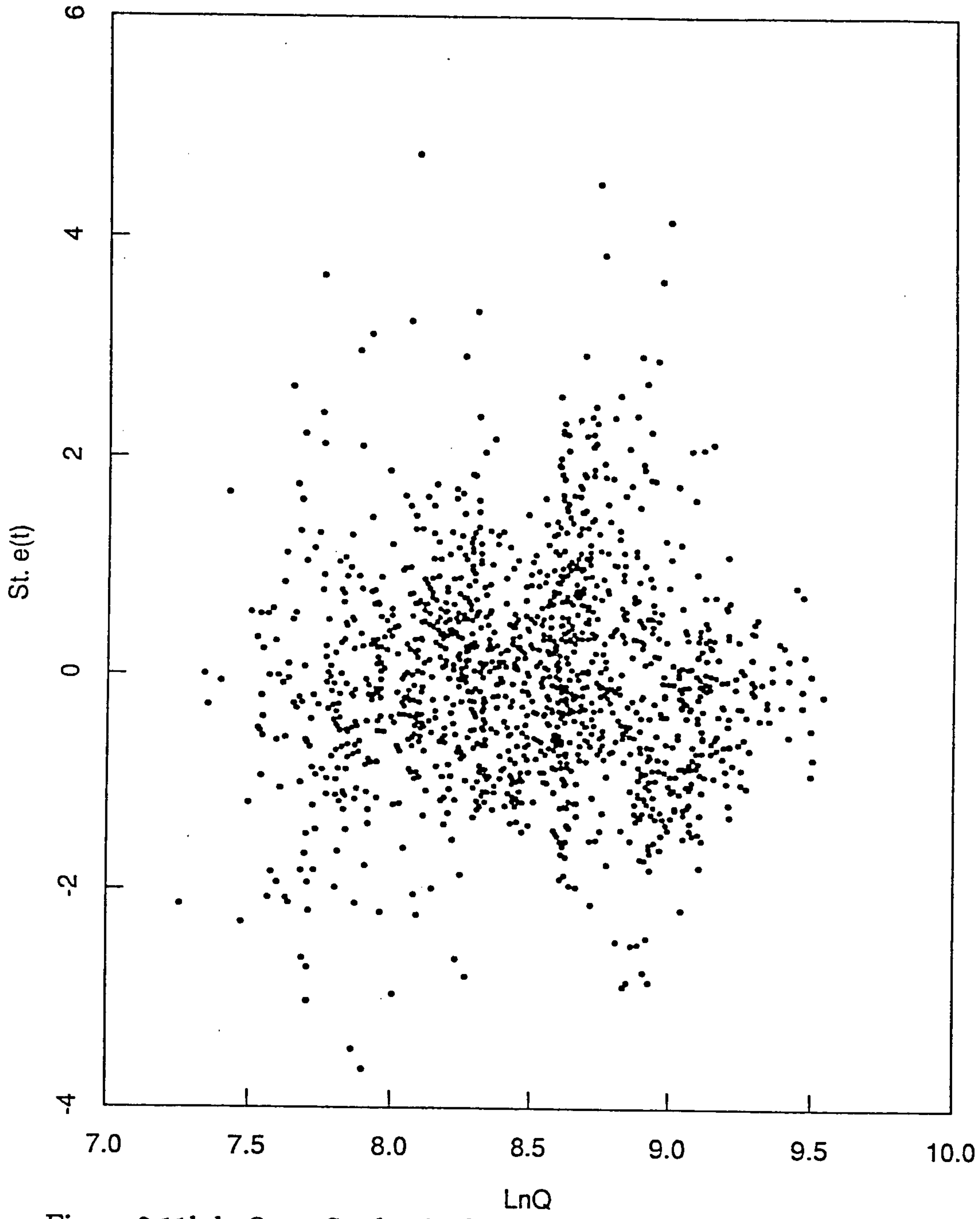

Figure 3.11b $\ln Q$ vs. Studentized residual for the falling stage of Mission.