STATISTICAL ANALYSIS OF THE TEMPORAL-SPATIAL STRUCTURE OF pH LEVELS FROM THE MAP3S/PCN MONITORING NETWORK

by

NHU DINH LE

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, 1986

Department of Statistics

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

Date August 20, 1986

ABSTRACT

The approach developed by Eynon-Switzer (1983) to analyze the spatial-temporal structure of a data set obtained from the EPRI monitoring network is applied to a data set obtained from the MAP3S/PCN monitoring network. In this approach, a spatio-temporal stochastic model, including deterministic components for seasonal variation and rainfall washout, is fitted to the data. The results indicate that the model fails to capture some of the features of the underlying structure.

In an effort to identify an appropriate model for the data, we examine the raw data in detail. An ANOVA model is fitted to the data. Different criteria such as Akaike, Schwarz, Mallows, etc, are used to identify the 'best' submodel (i.e. eliminate some terms in the full ANOVA model). The results indicate that it is possible to capture the deterministic component of the model with a much smaller model (i.e. fewer parameters). The normality of the residuals is also examined. The results indicate that the data from all stations except one can reasonably be approximated as coming from normal distributions.

TABLE OF CONTENTS

LIST OF TABLES

TABLE

LIST OF FIGURES

FIGURE

ACKNOWLEDGEMENTS

I would like to thank Professor A. J. Petkau for his guidance and assistance in producing this thesis. I would like to express my gratitude to Professor J. V. Zidek for his careful reading and helpful comments on this work. I am indebted to Drs. A. Olsen and C. Watson of Batelle Pacific Northwest Laboratory for making the MAP3S/PCN data set available to me. I also would like to express my appreciation to the UBC Acid Rain group, particularly Professors P. Guttorp and P. Sampson, for extensive input during the course of this work.

This work was partially supported by the Environmental Protection Agency through a co-operative research agreement with SIAM's Institute for Mathematics and Society (SIMS) and by the Natural Sciences and Engineering Research Council of Canada.

1) INTRODUCTION

1.1) Inferences for Spatial-Temporal Data

Given a data set (pH levels in this case) obtained from a set of monitoring sites, it is often necessary in environmetrics to make inferences about events at nonmonitored sites. The solution of this problem usually involves some interpolation techniques (i.e. finding the 'best' estimates for the data at nonmonitored sites). Kriging (Delhomme 1978) is a simple method for interpolation to obtain contour maps over the whole region which contains the monitoring sites. However, the spatial covariance function for the whole region is required in order to do kriging. The spatial covariance function can possibly be modelled by using the estimated values of spatial covariances between pairs of monitoring sites, which in turn depend on the spatial-temporal stochastic model used to explain the raw data. Successful use of this scheme for inferences concerning spatial-temporal data requires an appropriate spatial-temporal stochastic model for the data.

In chapter 2 of this thesis, the approach followed by Eynon and Switzer (1983) is applied to the data set obtained by the MAP3S/PCN monitoring network. The results indicate that the spatial-temporal stochastic model developed for the variability of pH levels in their paper fails to capture some of the features of this independent data set. However, for the sake of illustration, we carry the approach through to its logical conclusion and obtain by kriging, contour maps for this data set.

In an effort to identify a more appropriate spatial-temporal stochastic model for the data set, the underlying structure of the deterministic components of the data are carefully examined in chapter 3. The results show that there is some seasonal structure in the pH levels. The pH levels are not influenced by the rainfall volumes which show no seasonal structure of their own. A saturated ANOVA model is fitted to the pH levels and the importance of each component of the model is examined. The results suggest that the deterministic components of the pH levels can be captured by a simpler model than the saturated one. Many criteria for comparing different models are also described.

Many of the models and methods of analysis employed in environmetrics are based on assumptions of normality; therefore, the structure of the residual stochastic component is examined by simple normal probability plots. The results indicate that the data from all sites but one have residual variation which can be reasonably approximated as being normally distributed.

1.2) Data Set Description

The MAP3S/PCN network has nine monitoring stations which are located in the northeastern part of the United States. Data were collected for each day on which rainfall occurred from 1976 to 1982; one data record was collected per storm (i.e. event data). However, the stations started collecting data at different times; only data obtained at stations which were active for the entire year are considered (i.e. data obtained in the first year of operation, but with start-up time much later than January 1^{st} of that year are ignored). About five percent of the data are missing (i.e. there was a storm but no data were reported for the pH level or volume); missing data are ignored in this work. As in the Eynon-Switzer analysis, only (field) pH level and Volume measurements are used although the data set provides much more information (e.g. Conductivity, Sulfite, Sulfate, H^+,...). The variables considered are Time (t, t=0 on Jan 1^{st}, 1976 and the unit is a day), Volume (rainfall volume was obtained from a raingauge and measured in millimeters) and pH level. The station locations, the first active dates, and the ADS (Acid Deposition System) site identity are given in Table 1.1. Table 1.2 provides the number of rainfall events and the averages and standard deviations of the corresponding pH levels and rainfall volumes for each station in each year.

1.2.1) pH levels

The standard deviations of the pH levels are not entirely consistent from year to year and from station to station. The standard deviation of the data at station 020b (Illinois, Illinois) is consistently larger than that at other stations; this excess variability is easily seen in the plots of the raw data (Figures 1.1.a–1.1.d). Station 065a (Penn State, Pennsylvania) exhibits a large standard deviation (.61) in 1980 due to an outlying data-value (a pH level of 7.31 in December 1980; see Figure 1.1.d); the standard deviation calculated without this single pH level is .30, a value which is perfectly consistent with the standard deviations of other years at this station (Table 1.2).

The data corresponding to four stations (020b, 044a, 048a, 065a) are plotted in Figure 1.1 (pH levels against Time). These four plots do not demonstrate any obvious long-term relationship between pH levels and Time; of course, seasonal patterns may be obscured by the severely compacted Time axis. However, the plots do show differences in the data corresponding to different stations. Station 020b (Figure 1.1.a) has a large variability in its pH levels; station 048a (Figure 1.1.c) has distinctly smaller variability. Stations 044a and 065a (Figures 1.1.b and 1.1.d) display the same overall pattern of relatively small variability. One difference is that there are some possible outliers in the data obtained at station 065a. The plots of the data of other stations (not included in this report) are similar to one of the above figures.

1.2.2) Rainfall volumes

The average rainfall volumes vary considerably from year to year and from station to station (Table 1.2). The average rainfall volumes in 1979 are greater than those of other years at all stations but stations 044a and 048a. The standard deviations of the rainfall volumes are large (Table 1.2), and vary greatly from year to year and from station to station. At most stations the standard deviations in 1979 are exceptionally large relative to those of other years; for example, at station 013a, the standard deviation is 29.49 in 1979 while those in other years are less than 15.28. Rainfall volumes are plotted against time for each station separately to examine whether there is any obvious pattern in the data. The resulting plots do not suggest any obvious long-term relationship between the rainfall volumes and Time. Overall, all plots are similar; the plots of four stations (020b, 044a, 048a, 065a) are presented in Figure 1.2. The great variabilites in the rainfall volumes are clearly demonstrated in all plots. There is one possible outlier at station 020b in August of 1979; the rainfall volume is 157.14 mm (the rest are less than 90.00 mm) and the corresponding pH level is 4.22.

2) EYNON-SWITZER APPROACH

2.1) Introduction

One general objective of the study of rainfall acidity is the determination of 'best' estimates of a spatial process describing geographic fluctuations in the acid levels of rainfall at some unobserved locations using the observed data obtained at a limited number of stations monitoring rainfall acidity. These estimates would allow consideration of the optimal design of networks to monitor the long term and potentially damaging effects of the components of acid rain.

An approach for finding these estimates is provided by Eynon and Switzer (1983). In their paper, a parametric model is developed using data obtained from the Electric Power Research Institute (EPRI) network. A spatio-temporal stochastic model, including deterministic components for seasonal variation and rainfall washout and stochastic components for spatial, temporal, and measurement variation, is fitted to the data. Using the method known as Kriging, the best linear unbiased estimates (BLUE) of seasonal and rainfall adjusted yearly average pH over the monitoring region are obtained.

It should be noted that Eynon and Switzer use the data set obtained by the EPRI network which also has nine stations, each with two years of data. The geographical locations of stations in both networks are presented in Figure 2.1. Although the geographical locations of the stations in the two networks differ, both networks cover essentially the same area. One difference is that rainfall volume is measured in inches in the EPRI network.

In this chapter, we present the results of an attempt to validate the Eynon-Switzer model using the data obtained by the MAP3S/PCN monitoring network. The attempted validation follows the Eynon and Switzer paper in a step-by-step fashion. We also present some of the corresponding results from their paper using data obtained from the EPRI network.

2.2) The Eynon-Switzer Model

Eynon and Switzer propose the following model under which a single year's data are analyzed :

$$
\begin{aligned}
-\log_{10} H^{+}(x,t) &= pH(x,t) \\
&= \alpha(x) + \beta(t) + a(x)\sin\frac{2\pi t}{365} + b(x)\cos\frac{2\pi t}{365} \\
&\quad + \log_{10}\frac{c\cdot I(x,t)}{1-\exp\left[-c\cdot I(x,t)\right]} + \epsilon(x,t),
\end{aligned}
$$

where

x denotes the two-dimensional location under consideration,

t denotes the number of days starting from January 1^{st}, 1976,

$I(x,t)$ denotes the observed rainfall in millimeters at location x on day t (Eynon-Switzer used inches in their analysis),

$pH(x,t)$ denotes the observed rainfall pH at location x on day t, given $I(x,t) > 0$,

$H^{+}(x,t)$ denotes the hydrogen-ion concentration at location x on day t in gram-atoms per litre,

$\alpha(x)$ is a stationary autocorrelated spatial process describing the geographic fluctuations in annual pH level,

$\beta(t)$ is a zero-mean stationary autocorrelated time process describing temporal fluctuations not attributed to seasonal variation, independent of $\alpha(x)$,

$\epsilon(x,t)$ is a zero-mean stationary white-noise measurement error process, independent of $\alpha(x)$ and $\beta(t)$,

$a(x), b(x)$ are location-specific model constants which describe seasonal fluctuations, and

c is a washout rate constant.

The correction for rainfall volume (i.e. the logarithmic term in the pH model) represents the simplest scavenging process in which the number of hydrogen ions picked up per unit time per unit volume of rain is proportional to the remaining atmospheric concentration. This process produces lower acidity (higher pH) for larger rainfall volume. The limit of the logarithmic term is zero as the rainfall volume tends to zero. Since $\beta(t)$ and $\epsilon(x,t)$ are zero-mean residual processes, $\alpha(x)$ may be interpreted as the seasonally adjusted maximum acidity potential at location x.

2.3) Preliminary Fitting of The Model

In this section all parameters in the model are estimated by using the BMDP statistical computer package; the program used is Derivative Free Non-Linear Regression, P:AR. For each of the study years, the separate fitting of the data to the model proceeds in several stages. First, the constant coefficients $a(x), b(x), c(x)$ are fitted by non-linear least squares separately at each station location x_i, i=1,...,9. The model used for the data (corresponding to events at times t_j in the study year under consideration) being fitted in this step is:

$$pH(x_i, t_j) = a(x_i)\sin\frac{2\pi t_j}{365} + b(x_i)\cos\frac{2\pi t_j}{365} + \log_{10}\frac{c(x_i)\cdot I(x_i, t_j)}{1-\exp[-c(x_i)\cdot I(x_i, t_j)]}$$

$$+d(x_i) + \epsilon(x_i, t_j).$$

The values of the fitted parameters $\hat{a}(x_i), \hat{b}(x_i), \hat{c}(x_i), \hat{d}(x_i)$ for each year, and each station are presented in Table 2.1. Note that there are some negative values of $\hat{c}$'s at stations 020b and 057a. These values should always be positive in principle since c is defined as a washout rate. However, the magnitudes of these values are always very small. These negative values could possibly be the result of the round-off errors in the nonlinear fit of the model.

Figure 2.2 shows the phases and amplitudes of the nine fitted seasonal curves for each year separately. Amplitude is calculated as $\sqrt{\hat{a}^2 + \hat{b}^2}$, and phase is the value of t which minimizes $\hat{a}\sin\frac{2\pi t}{365} + \hat{b}\cos\frac{2\pi t}{365}$. Although most of the minimum points of the seasonal curves occur in the summer, the amplitudes and the times at which the minimum occurs for any station do not show great consistency from year to year. This fact is supported by the results presented in Tables 2.2.1 and 2.2.2 obtained by applying the ANOVA and the MEDIAN POLISHING decomposition techniques to the fitted coefficients $\hat{a}(x_i), \hat{b}(x_i)$. Results in the Eynon-Switzer paper suggests a similar problem. Two figures (similar to Figure 2.2) in their paper show the same inconsistency described above.

The results obtained by applying the techniques described above to the fitted coefficients $\hat{c}(x_i), \hat{d}(x_i)$ are also presented in Tables 2.2.3 and 2.2.4. Each table contains two parts. In Part (a) the average is used to do the decomposition. The coefficients are considered as the entries in a two-way table (year×station). Each entry could be represented as the sum of overall average, station effect, year effect and residual. The decomposition, then, is done by using an ANOVA-type approach; for example, the year (row) effect is taken to be the difference between the row average and the overall average. In Part (b) the Median Polishing approach is used to do the decomposition. Each entry in the year×station table is replaced by the difference between itself and the row median. The median of these row medians is taken to be the overall effect. The year effect is taken to be the difference between the row median and the overall effect. The station effect is taken to be the column median of the new table. The residual is then the difference between the entry of the new table and the station effect; the sum of these effects and the residual results in the original entry of the two-way table.

In Tables 2.2.1 and 2.2.2 of estimated values of the seasonal coefficients $a(x)$ and $b(x)$, there are many large residuals. These values appear at different stations and different years without any apparent pattern. Some year effects are much larger, in magnitude, than others. This suggests a considerable interyear variation in the year effect The station effect behaves the same way. Therefore, we may infer that the periodic coefficients $a(x)$ and $b(x)$ are inconsistent from year to year, and from station to station. Note that if the proposed model fitted the data well enough, then one would expect the periodic coefficients to be consistent at least from year to year. This suggests that these data do not have the periodic behaviour postulated in the model. We also apply the above decomposition techniques to the year×station tables of phases and amplitudes. The results which are not included in this thesis support our earlier conclusion that the phases and amplitudes are somewhat inconsistent from year to year.

It is not surprising that the estimated values of $d(x)$ (decomposed in Table 2.2.4) show little consistency from year to year, and from station to station. The overall effect, $d(x)$, should be different from station to station and possibly also from year to year.

Table 2.2.3 shows some consistency in the estimated values of c.

There is only one large residual (at station 065a, in 1980) which is greater (but not very much) than the other residuals. The magnitudes of the year effect are approximately the same. The station effect, however, shows some variation. There are two relatively large values of the station effect which correspond to stations 020b and 048a. This points to possible differences in station effects. Besides, some of the residuals corresponding to stations 048a and 065a vary considerably from year to year (and also from station to station). Therefore, estimates of a common washout rate c obtained by using the data from all of the stations in each year separately, may not be consistent from year to year in conflict with a suggestion in the Eynon-Switzer paper; this is now examined in more detail.

Following Eynon and Switzer, we fit a common value of the washout rate parameter c across all stations for each year separately. Using the previous estimates of $a(x_i), b(x_i), d(x_i)$ (different for each station), we fit a common washout rate c for each year separately. The fitted values of the parameter c are:

Year	1977	1978	1979	1980	1981	1982
$\hat{c}$	.0405	.0528	.0189	.0555	.0304	.0489
se($\hat{c}$)	.0050	.0060	.0037	.0067	.0095	.0049.

Here se($\hat{c}$) denotes the nominal asymptotic-theory standard error of $\hat{c}$, based on independence assumptions on the residuals. As a rough check for significant year to year differences in these c-values, we apply the Newman-Keuls procedure. The c-values are assumed to be independent, with a common standard error of approximately .006 (this value is the root-mean square of the standard errors displayed above). While the assumptions giving rise to this procedure are not satisfied exactly, nevertheless, the results give us some rough indication of the behaviour of these c-values. As expected, there are some statistically significant differences (at the five percent level). Specifically, the c-value obtained for 1979 is significantly different from the c-values obtained for 1977, 1978, 1980, 1982; and the c-value of 1980 is also significantly different from the c-value of 1981.

Note: The fitted values of the common washout rate parameter c given by Eynon-Switzer are:

$$\begin{aligned} \hat{c} &= 0.92 \text{ for year 1 (Sept. 78 - Aug. 79)}, \\ &= 1.10 \text{ for year 2 (Sept. 79 - Aug. 80)}. \end{aligned}$$

These values correspond to rainfall volume measured in inches. To make these values comparable with our values of $\hat{c}$ (corresponding to volume measured in millimeters), we divide $\hat{c}$ by 25.4 (1 inch = 25.4

millimeters). The converted values of Eynon-Switzer's $\hat{c}$ are respectively $\hat{c} = 0.92/25.4 = .0362$, and $\hat{c} = 1.10/25.4 = .0433$.

These values of $\hat{c}$ agree with ours. Specifically, they are about in the middle of our range of values of $\hat{c}$.

The combined corrections for rainfall volume (using a common washout rate $\hat{c}$ for each year) and seasonality reduce the overall variation of an annual pH series; the residual standard deviation of a series is reduced by 0 - 30 percent from that of the corresponding uncorrected series. The ratios of sums of squares of residuals for the corrected series to that of the uncorrected series are presented in Table 2.3. In some cases, essentially no reduction is obtained by using the combined corrections. The reductions corresponding to station 020b in different years are always very small. The results presented in Table 2.3 do not show any obvious relationship between the percentage of the reduction and the stations or the years.

We note that according to Eynon and Switzer, these corrections produce a 0 - 20 percent reduction of the residual standard deviation with their data. In addition, they note that each correction separately contributes about one-half of the reduction. To see whether this is also true for our data, we fit the full model and the reduced model in which the volume correction is omitted (i.e. with only seasonal terms) to eight different data sets. Each data set corresponds to one year at a specific station. The results are presented in Table 2.4. The reduction in the residual sum of squares due to the seasonal correction is much higher in some cases and lower in others than the reduction due to the volume correction. It is not obvious here whether each correction separately contributes about one-half of the reduction.

2.4) Spatial and Temporal Components

With the estimated parameters $\hat{a}(x_i), \hat{b}(x_i)$ for each station, each year and the estimated common washout rate $\hat{c}$ for each year, the residual for the station at location x_i for collection day t_j is given by:

$$\widetilde{pH}(x_i, t_j) = pH(x_i, t_j) - \hat{a}(x_i) \sin \frac{2\pi t_j}{365} - \hat{b}(x_i) \cos \frac{2\pi t_j}{365}$$

$$- \log_{10} \frac{\hat{c}(x_i) \cdot I(x_i, t_j)}{1 - \exp[-\hat{c}(x_i) \cdot I(x_i, t_j)]}.$$

These $\widetilde{pH}(x_i, t_j)$ values are treated as estimates of the combined random components $\alpha(x_i) + \beta(t_j) + \epsilon(x_i, t_j)$.

Define:

$V_\alpha(u) = E[\alpha(x) - \alpha(x+u)]^2/2$: the variogram of the spatial process, and

$V_\beta(w) = E[\beta(t) - \beta(t+w)]^2/2$: the variogram of the temporal process.

These quantities describe the geographic and temporal variation of daily 'corrected' pH readings. Eynon-Switzer indicate that for convenience, the following two simple parametric models for the variograms were fitted:

$$V_\alpha(u) = \frac{h_1}{1 + [u' h_2 u]^{-1}},$$
$$V_\beta(w) = \frac{g_1}{1 + [w^2/g_2]^{-1}},$$

where g_1, h_1, g_2 are non-negative constants and h_2 is a symmetric positive definite 2×2 matrix.

The contours of the spatial variogram $V_\alpha(u)$ are concentric ellipses centered at the origin. The models satisfy the constraints of positive definiteness required of all variograms. The above variogram models are also used to fit our data as described in what follows.

2.4.1) Fitting the temporal variogram:

The corrected pH values, $\widetilde{pH}(x_i, t_j)$, are used to estimate an unsmoothed variogram;

$1/2[\widetilde{pH}(x_i, t_j) - \widetilde{pH}(x_i, t_{j'})]^2$ estimates $V_\beta(w) + \sigma_\epsilon^2$, for $w = t_j - t_{j'}$, where $\sigma_\epsilon^2 = var(\epsilon(x_i, t_j))$.

Note that Eynon and Switzer estimate σ_ϵ^2 by using duplicate pH readings. In fact, the replication variability was estimated separately at each station for each year. There were modest differences among these estimates but not so large as to warrant modelling σ_ϵ^2 as a function of location. In any event, their estimate, $\hat{\sigma}_\epsilon^2$, turns out to be small relative to temporal and spatial variability, and a common value of $\hat{\sigma}_\epsilon^2 = .019$ was obtained with their data. In our data, since no replicate observations are available, σ_ϵ^2 is absorbed into the temporal variogram. So $1/2[\widetilde{pH}(x_i, t_j) - \widetilde{pH}(x_i, t_{j'})]^2$ estimates $V_\beta(w)$.

All rainfall days separated by a lag w (at a fixed station, separately in each year) are used to estimate $V_\beta(w)$. The scatter plot of the estimates of the unsmoothed temporal variogram at station 072a in 1982 is presented in Figure 2.3. All possible estimates are plotted; there are 1653 data points. Many of these estimates are near zero, but the variability of the temporal variogram estimates is very large. This pattern is, however, a typical one for all other plots of the temporal variogram estimates at different stations.

Eynon and Switzer observe that the fitted value of the constant g_2

could not be distinguished from zero with their data. To see whether this is true with our data, we fit the variogram model $\frac{g_1}{1+(w^2/g_2)^{-1}}$ to four different data sets corresponding to different combinations of year and station; the results are presented in Table 2.5. Our results support the conclusion of Eynon and Switzer (that the nominal asymptotic standard error of the estimate is always bigger than the estimated value). We would propose to let $g_2 = 0$.

It may be inferred that the temporal variogram has a constant value g_1. Using the data from all years and all stations, the estimated value of g_1 is $\hat{g}_1 = .1287$. In Table 2.6, different constants g_1 which are estimated using the data from each year and each station separately are presented. Applying the previous decomposition techniques to these values, we obtain the results presented in Table 2.7. There are some unreasonably large values of residuals corresponding to station 020b. The effect of station 020b is 0.31. This is very large relative to the other values whose magnitudes do not exceed 0.08. This suggests an unusual pattern in the data obtained at station 020b; recall the excessive variability of the data at station 020b noted earlier. Specifically, the values of $\hat{g}_1 = 0.805$ in 1979 at station 020b is very large relative to all others. Except for these consistently large values of station 020b, the results do not show any obvious pattern between the remaining values.

Note: Our value of $\hat{g}_1$ of .1287 is bigger than the value of $\hat{g}_1 + \hat{\sigma}_\epsilon^2 = .079 + .019 = .098$ obtained by Eynon-Switzer. One possible reason for this is that we appear to have outliers in our data. The excessive variability of the data at station 020b might be another possible reason.

2.4.2) Fitting the spatial variogram:

The corrected pH values, $\widetilde{pH}(x_i, t_j)$ are used as follows to obtain the unsmoothed variogram:

$$1/2[\widetilde{pH}(x_i, t_j) - \widetilde{pH}(x_{i'}, t_j)]^2 \text{ estimates } V_\alpha(u) + V_\beta(0^+),$$

where u is the vector joining station pairs $(x_i, x_{i'})$. $V_\beta(0^+)$ comes into the expression because the time unit was previously chosen as one day, and two readings of pH levels at two stations on a same day (i.e. same t_j) might not have the identical collecting times so that the time lag is always 0^+. Hence

$$1/2[\widetilde{pH}(x_i, t_j) - \widetilde{pH}(x_{i'}, t_j)]^2 - \hat{g}_1 \text{ estimates } V_\alpha(u).$$

Of course, $V_\alpha(u)$ can be estimated directly only for arguments u corresponding to two stations with events recorded on the same day.

With our data set, there are 2848 estimated values for the unsmoothed spatial variogram corresponding to two-dimensional vectors u joining the 36 station pairs. An attempt to fit $V_\alpha(u)$ to this data set failed because the algorithm failed to converge.

We note that Eynon and Switzer do not report any difficulties in fitting the spatial variogram.

To find the estimated values of h_1 and h_2 for our data set, we try another approach. For each station pair, the simple average of all the available estimates of the unsmoothed spatial variogram is considered as a new data point. The number of unsmoothed estimates entering this average is treated as a case weight. This new data set has 36 data points, with a case weight attached to each point. However, the attempt to fit $V_\alpha(u)$ to this new data set also fails; each of three different algorithms, Gauss-Newton and Marquardt algorithms (SAS statistical computer package), and Derivative-Free Non Linear Regression (BMDP package), fails to converge to a solution.

The plot of this new data set is in Figure 2.4. The plot, however, shows that the proposed spatial variogram model is totally inappropriate for our data set. There are many negative estimates possibly because of the large estimated value of $V_\beta(0^+)$ (i.e. $\hat{g}_1 = .1287$). And there are also some unreasonably large values of the unsmoothed spatial variogram estimates. These estimates are always from those station pairs where one of the two stations was station 020b (Illinois, Illinois). The difficulties encountered in this fitting may be due to these values. This is another indication of the atypical behaviour of the data obtained at station 020b. The plots of the unsmoothed spatial variogram estimates for each year separately (not included in this report) have approximately the same pattern as the above.

The final step in the Eynon-Switzer approach is the interpolation step, called kriging. To carry out this step, an estimated spatial variogram model is required. Since there have been indications that station 020b could be the source of many problems that we have encountered in applying the approach to this data set, we will repeat the previous steps of the approach to the data set with station 020b removed. This may allow us to find estimates of the parameters h_1 and h_2 in the spatial variogram model used by Eynon and Switzer. Besides, we would like to find out how well this approach work if we remove station 020b. This will be done in the next section.

2.5) Fitting Without Station 020b

In this section we examine whether we can obtain estimates of the parameters h_1 and h_2 from the data with station 020b removed. We repeat the previous steps of the Eynon-Switzer approach on the reduced data set and present the results in what follows.

The estimates of parameters a,b,c,d in the Eynon-Switzer model are the same as before. Now we use these estimates to fit a common c value for each year separately. The fitted values of the parameter c are:

Year	1977	1978	1979	1980	1981	1982
$\hat{c}$	.0405	.0622	.0297	.0654	.0492	.0530
old $\hat{c}$	.0405	.0528	.0189	.0555	.0304	.0489
se($\hat{c}$)	.0050	.0054	.0033	.0060	.0044	.0049

where the notation is the same as in section 2.3. The estimated values of c for the years in which the data from station 020 are removed turnout to be larger than those obtained earlier; the largest increase, 0.0188, is in 1981. As before, we apply the Newman-Keuls procedure to check for significant year to year differences in these c-values. An approximate common standard error of $\hat{c}$ is about 0.005. The result obtained in this step is generally similar to that in section 2.3 although there are differences. Specifically, the c-value for 1980 is no longer significantly different from the c-value of 1981, and the c-value obtained in 1979 is significantly different from all the rest.

We use estimated values of a, b and a common c to estimate the temporal variogram V_β as in section 2.4.1. As before, the fitted value of g_2 is negigible (i.e. the temporal variogram would have a constant value g_1). Using all the data together (excluding station 020b), the estimated value of g_1 is 0.095; this value is comparable to the Eynon-Switzer estimated value of $\hat{g}_1 + \hat{\sigma}_\epsilon^2 = 0.098$. This supports our claim that our earlier estimated value of g_1 is larger than Eynon and Switzer's because of the excessive variability of the data obtained at station 020b.

Using this $\hat{g}_1$, we obtain the unsmoothed estimates of the spatial variogram as in section 2.4.2. Our attempt to fit the model for V_α to these estimates fails (i.e. the nonlinear regression fit fails to converge to a solution for h_1 and h_2) using all the different approaches described in section 2.4.2.

To examine the unsmoothed estimates of the spatial variogram, we plot them in Figure 2.5. The points with unusually large values in Figure 2.4 do not appear in this plot; however, the proposed model for the spatial variogram still does not seem to be appropriate. Note that there are still many estimated values of the spatial variogram which are negative. This suggests that we may still be overestimating the temporal variogram g_1, although our estimated value of g_1 is now comparable to Eynon and Switzer's.

So the removal of station 020b from the data set has not resolved the difficulty in obtaining estimates of the parameters h_1 and h_2 in the Eynon-Switzer spatial variogram model. Consequently, the interpolation procedure called kriging, undertaken by Eynon-Switzer could not be carried out with this spatial variogram. However, to demonstrate the complete approach of Eynon-Switzer, we consider an alternate spatial variogram model in the next section.

2.6) Kriging With a Different Spatial Variogram

We have learned from our earlier analysis that the Eynon-Switzer model does not appear to be appropriate for this data set. Consequently, the interpolation step based on the residuals from this model seems inappropriate. However, to demonstrate the considerable value of the Eynon-Switzer approach in analyzing spatial-temporal data (when the model is appropriate), we carry out the interpolation step with the spatial variogram

$$V_{\alpha}(u) = a \cdot I_{(u \neq 0)} + b \cdot ||u||,$$

where u is the vector joining station pairs; $||u||$ is the length of u and $I_{(u \neq 0)}$ is the indicator function. The parameters a and b are fitted with constraints $a \geq 0, b \geq 0$, using the above unsmoothed estimates of the spatial variogram. The estimated values of a and b are 0.000 and 0.005 respectively. For each year, the estimated values of the realized spatial process $\alpha(x)$ at the station locations $x = x_i$ can be obtained by

$$\hat{\alpha}(x_i) = \frac{\sum_j \widetilde{pH}(x_i, t_j)}{n_i},$$

where n_i is the number of rainfall days at station i in the given year. The estimates $\hat{\alpha}(x_i)$ for each year are presented in Table 2.8. Because of the rainfall volume correction, these estimated values are smaller than the simple averages of the uncorrected pH levels (station 020b in 1979 is an exception; this may be due to the excessive variability of pH levels at this station 020b in 1979). At other stations, the differences range from 0.01 to 0.20 and they vary from year to year and from station to station. To put the corrected pH values of Table 2.8 on the original scale we would need to add a number of pH units corresponding to the washout associated with the average daily rainfall at that station in that year. Although the washout rate c is assumed to be common across stations in each year, the large differences in average daily rainfall across stations in each year mean that these adjustments will differ somewhat from station to station in each year. For example, in 1979 when the estimate of the washout rate is $\hat{c} = .0189$, these adjustments range from 0.05 to 0.10. Similarly, in 1982 when $\hat{c} = .0489$, the adjustments range from 0.08 to 0.22. It is important to add these adjustments to the corrected pH levels before doing the interpolation because estimates at unobserved locations (i.e. contour maps) will be substantially influenced by these values.

Since the spatial-temporal model for the pH levels does not seem to be appropriate, resulting contour maps may be misleading. We would like to demonstrate how the kriging method works, but do not intend to take the interpolating results seriously. For this reason adjusting the corrected pH levels seems unneccessary. The interpolation step described below is carried out without first adjusting the corrected pH values.

To obtain estimates of the pH levels for unobserved locations (subsequently, contour maps), we use the interpolation technique called kriging (Delhomme 1978). The 'best' linear unbiased estimate of $\alpha(x_0)$ at an unobserved location x_0 is given by

$$\alpha^*(x_0) = \sum_{i=1}^{9} \lambda_i(x_0) \cdot \hat{\alpha}(x_i),$$

where the weights $\lambda_1, ..., \lambda_9$ are chosen to minimize

$$v(x_0) = E[\alpha^*(x_0) - \alpha(x_0)]^2$$

subject to $\sum \lambda_i(x_0) = 1$. With stationarity assumptions about the process underlying the data (c.f. Journel and Huijbregts (1978)), the quantity $v(x_0)$ can be expressed as a quadratic form in the λ_i's whose coefficients depend on the spatial variogram V_α, the temporal variance parameter g_1, and σ_ϵ^2. We minimized this quadratic form for each x_0 on a fine grid over the eastern United States using the above estimated parameters for the spatial variogram. The contour maps of these resulting interpolated values are presented in Figure 2.6 (produced using the computer language S) for each of the four years from 1979 to 1982. The contour maps could possibly be useful in investigating the pattern of changes from year to year. However, as mentioned earlier, these results could be misleading, so attempting to draw conclusions from these results seems unwise.

2.7) Conclusion

In this attempt to validate the Eynon-Switzer model using the MAP3S/PCN data we obtain some results which are consistent with those based on the EPRI data set. However, some of the results differ. Specifically, there are significant differences between the estimated values of the yearly washout rate c. Our estimated constant value of the temporal variogram V_β is larger than that reported by Eynon and Switzer, possibly because of some potential outliers in our data set (including station 020b). This is supported by the results obtained in section 2.5 using the data set with station 020b removed, where the estimated value of V_β is about the same as that obtained by Eynon and Switzer (.095 vs .098). The spatial variogram model proposed by Eynon-Switzer does not appear to be appropriate for the MAP3S/PCN network, possibly because of the differences in geographical location between the two networks. Overall, the indication obtained from our analysis is that the Eynon and Switzer model does not completely capture the structure of this data set. However, to investigate their complete approach throughly, we use an alternate spatial variogram model to obtain the contour maps for 1979 to 1982. But as e we have already concluded that the proposed spatial-temporal model for the pH levels appears to be inappropriate, drawing conclusions from these contour maps seems ill-advised.

Since this approach is unsuccessful for this data set, we propose to examine the raw data in more detail with a view towards identifying structure which may be present. This will be done in the next chapter.

3) DATA EXPLORATION

We now start a careful examination of the data in an attempt to uncover any underlying structure which may be present. We examine a number of factors which may be expected to influence the pH readings and investigate the possibility of consistent relationships between pH levels and these factors. There may also be some relationships between these factors.

3.1) Factors to Be Examined

As described earlier, there are about 3000 data points in the data set. Each data record contains the time of the event (the storm), the location of the station, the rainfall volume of the storm, and other information. Note that in this examination one pH reading of 7.31 at station 065a in 1980 is removed. The new mean and standard deviation for this year at this station are 4.04 and 0.33 respectively, while the old ones are 4.14 and 0.61 (Table 1.2). The removal of this obvious outlier makes the new standard deviation of this year consistent with other years at this station (about 0.3).

The location of the station should be one important factor. Most stations are located in heavily industrialized areas; therefore, the pollutants in the sky at each station are possibly influenced by the chemicals released from the factories in that area. There are also possibly some long-range transportation effects, that is tall remote smokestacks emissions are carried by strong winds at high altitudes over very long distances to the location. These contaminants may affect the acidity of the storm at the station under consideration. For these reasons, we expect the effects on the pH levels of the geographical locations of stations may be different from station to station. In our study, the station factor is a categorical variable with nine categories; each specifies a single station.

Another important factor is the time of the storm. In some factories, the volume of production is seasonal (i.e. for some specific periods, production is very high, and in some other periods, production is low). During periods of high volume production, more contaminants are released into the sky. These substances may affect the pH reading of the storm. Further, during the summer season, forest fires may pollute the sky; these obviously change the chemistry of the storm, which will in turn affect the pH readings of that storm. For these reasons, we expect the time (i.e. seasonal) factor is important for this study. To account for seasonality, time is divided into years and months. Note that Egbert and Lettenmaier (1986) divided time into years and seasons (3 or 4 months each), but did not discover any obvious patterns in pH levels. For this reason, we use a shorter time period (i.e. one month) in this analysis.

Another important factor is rainfall volume. It is generally believed that the pH reading varies monotonically with the rainfall volume. However, in the previous chapter, our observations of the effect of volume as represented by the scavenging term used by Eynon-Switzer (1983) do not support this belief. Therefore, we will examine more general relationships between rainfall volumes and pH levels.

3.2) Preliminary Examination of The Structure of The Data

We first examine the relationships among Time, Volume, Station, and also between these factors and the pH levels. In this section we use monthly averages of pH levels and Volumes instead of individual values. The monthly average pH level (or Volume) is just the simple average over the events in each month. Since these monthly averages smooth the original data (i.e. reduce the variability), relationships may be more clearly exhibited in the averaged data than in the original (individual) data. If there are no obvious relationships in the averaged data, relationships in the original data set seem unlikely. One potential problem with this approach is that taking averages could cancel out some effects which we may be able to detect in the individual data. However, this is unlikely to occur since with only a few observations available in each month (about 1-8 points), effects are very hard to detect even if they are present in the individual data. Note that relationships suggested by the monthly-average data can always be examined subsequently in the individual data. Working with the average data also has the practical advantages of reducing the size of the data set to about 500 data points from about 3000 data points. In the following sections, we examine the relationships among the factors in detail using the averaged data set.

3.2.1) Relationships between pH and Month, pH and Station

We would like to find out whether or not the Station and Time factors affect the readings of pH levels. The Time factor has two components: Year and Month. In each year, there are 12 monthly averages at each station. To examine the relationships among these factors, we will look at the data of each year (all stations) separately. If there are obvious relationships in each year, we hope to be able to recognize how they change over the years. This would subsequently help us to understand the relationships among the factors for all years of data.

The monthly average data in each year are fitted with the following additive model:

$$p\bar{H}_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij},$$

where

$p\bar{H}_{ij}$ denotes the pH average of Month j at Station i,

μ denotes the overall effect,

α_i denotes the Station effects, i = 1 to 9,

β_j denotes the Month effects, j = 1 to 12,

ϵ_{ij} denotes the residuals.

The residuals may contain not only the pure errors, but possibly also some other effects such as a rainfall volume. We proceed to obtain rough estimates of the parameters by using ANOVA-like and median polish decompositions which are described in detail in section 2.4. The monthly averages are arranged in a 2-way table; the factors are Month and Station. We fit the model by this rough method for years 1977 to 1982 separately. The estimated values of the parameters α and β for the years from 1977 to 1982 are presented in Tables 3.1 and 3.2 respectively.

The estimates of μ are:

Year	1977	1978	1979	1980	1981	1982
ANOVA	4.14	4.10	4.27	4.18	4.14	4.21
Median	4.14	4.07	4.22	4.14	4.16	4.25

Note that similar estimates of μ are obtained by both methods. The monthly averages range only from about 3.5 to 5.5, and most of them are about 4. Therefore, the values of the overall effect obtained by taking the overall average and the median should be very similar. The differences of the estimates from year to year are not large. If we had many more years of data, then we could investigate whether there is a true yearly effect or a trend (pattern) over the years by examining the estimated values of μ. Unfortunately, these issues could not be examined thoroughly because we have only six years of data. Although the six estimated values of μ (i.e. for the six years) are similar in magnitude, it is not reasonable to ignore the yearly effect on the pH levels. Therefore, we still assume that there is some yearly effect on the pH levels.

The station effects ($\hat{\alpha}$) for the different years are plotted against Station in Figures 3.1.1 (ANOVA) and 3.1.2 (Median Polish) to see if there is any consistent pattern from year to year. Similarly, the monthly effects ($\hat{\beta}$) for the different years are plotted against Month in Figures 3.2.1 and 3.2.2. Note that the plots for the two methods of decomposition are qualitatively similar for both sets of estimates.

Figure 3.1 shows that the station effects are quite different from station to station. In Figure 3.1.2, there are large positive effects from stations 013a (0.05-0.27) and 020b (0.12-0.34). There are also some large negative effects from station 057a; for example, the effect is -0.30 in 1979. The effects for the remaining stations are very small or centered around zero. The changes in the station effects from year to year are considerable for most stations; stations 048a and 072a are exceptions. In general, Figure 3.1.1 has similar features, although there are some differences. The effects do not change from year to year at station 013a. The effects vary more from year to year in Figure 3.1.1 than Figure 3.1.2. Overall, the effects are quite different from station to station in both plots. This observation supports our belief that the geographical location of a station is an important factor in explaining the pH levels obtained at that station.

Note that at station 020b, both estimates of the effect of Station for 1979 are unusually large (Figure 3.1). The reason is that in November 1979, there was only one storm with an unusually high pH level of 7.12;

therefore, the monthly average of this month is greater than those of other months (7.12 compared to about 4.0-5.5). Moreover, the estimate for 1979 (Figure 3.1.1) based on the ANOVA decomposition is greater than that from the Median decomposition (Figure 3.1.2). This may be explained by the robustness of the Median Polish decomposition. The monthly average of November in 1979 is obviously an outlier. The decomposition using medians reduces the effect of the outlier on the estimated values of the parameters more than the decomposition using averages (ANOVA).

Figure 3.2 shows some interesting results. The monthly effects vary considerably from year to year. However, the effects change in a somewhat consistent manner from month to month. Generally, the effects decrease slowly from January to August, increase sharply from September to November, and then decrease a little in December. Note that this general pattern only becomes apparent when the effects for all years are plotted together. In each year, the pattern of the effects is somewhat different from the general pattern. For example, the effect increases from January to February (sharply) in 1981 and from March to April in 1982; the general pattern suggests a decrease in both cases. These plots also indicate that the minimum of the monthly effects always occurs in the summer. This result supports the expectation that levels of acidity increase in the summer. Overall, although there are differences in the patterns of different years, the general picture described above is the only apparent structure in the data.

3.2.2) Relationships between Volume and Station, Volume and Month

The monthly average of the pH levels adjusted for these additive station, monthly, and yearly effects, denoted by $\widetilde{pH}_{ij}$, is

$$\widetilde{pH}_{ij} = p\bar{H}_{ij} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j.$$

The results which follow are based on the adjusted pH levels obtained by using the estimated values of μ, α, β provided by the median polish decomposition. We use the median polish estimates here because there are possible outliers in the data set. The effect of outliers on the estimated values of parameters is much less pronounced with the median decomposition than the ANOVA method; therefore, the median decomposition is more appropriate.

Since we wish to determine the relationship between pH levels and Volumes, we proceed to find the station, monthly, and yearly effects on the rainfall volumes so that we can adjust the rainfall volume for these effects as we did for the pH levels in the previous section. This is done in what follows.

The relationship between Volume and Station and Month can be examined by an approach similar to that in section 3.2.1. In each year we fit the following model:

$$\bar{V}_{ij} = \mu + \gamma_i + \omega_j + \epsilon_{ij}$$

where

$\bar{V}_{ij}$ denotes the volume average in Month j at Station i,

μ denotes the overall effect,

γ_i denotes the Station effects, i = 1 to 9,

ω_j denotes the Month effects, j = 1 to 12,

ϵ_{ij} denotes the residuals.

We obtain estimates of the parameters μ, γ, and ω by using the median polish and ANOVA-like decompositions. We fit this model to the data of each year separately from 1977 to 1982. The estimated values of μ are:

Year	1977	1978	1979	1980	1981	1982
ANOVA	14.42	12.82	17.07	10.53	10.77	12.00
Median	15.32	14.03	20.31	11.00	12.39	13.43

The estimates of μ are quite different from year to year. Therefore, we infer that there is a substantial yearly effect on the average volumes. The distribution of the monthly average rainfall volumes has a long left tail in each year (all stations); consequently, the median is greater than the average. This is reflected in the above table where the estimates of μ based on the median decomposition are in every case larger than those based on the ANOVA decomposition.

The estimated values of the parameters γ and ω are presented in Tables 3.3 and 3.4, respectively and are plotted in Figures 3.3 and 3.4 respectively. Figure 3.3.1 suggests some differences between stations. The effects vary greatly from year to year at some stations. Specifically, the effect for 1979 at station 013a is much greater than that for other years. Also, the effect for 1979 at station 048a is less pronounced than that for other years. The difference between 1981 and 1982 at station 171b is large (about 8). Figure 3.3.2 suggests a qualitatively similar general pattern although there are some differences. The effect of 1979 at station 020b is very small in Figure 3.3.2 but it is large relative to other years in Figure 3.3.1. This is due possibly to the fact that the average volumes are greatly different from month to month in 1979 at this station. As a result, the average and the median of the monthly rainfall volume averages of this year (station 020b) are greatly different; the average is about 20.71 mm and the median is about 7.82 mm. Consequently, the results of the two decomposition methods are quite different for year 1979 and station 020b. Note that there was no rainfall event in February 1979 at this station. Overall, Figure 3.3 suggests that the monthly volumes are substantially different from station to station.

Figures 3.4.1 and 3.4.2 show that the monthly effects do not exhibit any obvious pattern. Although the estimates are quite different from year to year, they are scattered around zero. This suggests that seasonality is not important for monthly average rainfall volumes. This result seems surprising; average rainfall volumes might be expected to differ from season to season. Since seasonality was important for pH levels and might be expected to be important for rainfall volumes, this issue is now examined further.

We apply the SUPER SMOOTHER developed by Friedman and Stuetzle (1982), to all individual (event) data. The Super Smoother uses local linear fits with varying window width determined by local cross-validation. In general, this smoother takes bivariate data and produces a smooth fitted relationship between the two variables. In this exercise we use Volume and Time as a bivariate data set. Both Volume and Time are measured as in chapter 2 (i.e. Volume is measured in millimeters and Time is the number of days from January 1st,1976 to the day of the event). The relationship produced by the smoother is Volume as essentially a constant over Time. The constant is 13.6 mm which is very close to the average of all individual rainfall volumes. We also apply the same procedure with Volume adjusted for the Year and Station effects. The result indicates that there is no obvious relationship between the adjusted Volume and Time (i.e. Month). The smoothed curve provided by the smoother does not resemble any standard functional form (Eynon-Switzer suggested a logarithmic relationship between Volume and Time).

These results suggest that the rainfall volume of a storm is not influenced by the time of the storm (i.e. there is no seasonality). It appears that the effect of the factor, Month, on the average rainfall volumes can be ignored.

The model for the monthly average rainfall volume can now be written as:

$$\bar{V}_{ij} = \mu + \gamma_i + \epsilon_{ij},$$

where everything is as previously defined. Using the estimated values of the parameters μ and γ, we obtain the adjusted average rainfall volume as follows:

$$\tilde{V}_{ij} = \bar{V}_{ij} - \hat{\mu} - \hat{\gamma}_i.$$

$\tilde{V}$ represents the monthly average rainfall volume adjusted by the yearly average and the station effect. Note that we use the median estimates to estimate μ and γ.

3.2.3) Relationship between pH and Volume

We now search for a relationship between monthly average pH levels and rainfall volumes using the adjusted data ($\tilde{V}$, $\widetilde{pH}$) instead of the original data ($\bar{V}$, $\bar{pH}$). Note that if there is indeed a relationship between pH and Volume then it would be more easily detected using the adjusted data than the original data. The effects of each factor (Year, Station, Month) on the volume and the pH level are different; therefore, under the influence of these different effects, the original data may suggest some relationship which is not true.

As a first step, $\widetilde{pH}$ is plotted against $\tilde{V}$ in Figure 3.5. Only data from 1979 to 1982 is used because only three stations were active in 1977 and only five stations were active in 1978. Station 171b is not included since it has only two years of data. There are about 400 data points. The figure shows that most data points lie near the origin (i.e. $\tilde{V} = 0$, $\widetilde{pH} = 0$). There are some unusually large values of $\widetilde{pH}$ and $\tilde{V}$. There are four large values of $\widetilde{pH}$ corresponding to the values of $\tilde{V}$ ranging from -10 to 5. There are also four large values of $\tilde{V}$ corresponding to $\widetilde{pH} \simeq 0$. With these points deleted, the figure does not show any irregularities. To examine these unusual points further, we plot the data from each station separately in Figures 3.6.1 to 3.6.8. These figures reveal that all of the four large values of $\widetilde{pH}$ and the largest value of $\tilde{V}$ are from station 020b. They also show that the other three large values of $\tilde{V}$ are from some possible outliers in the data. Indeed, at station 013a, there is one monthly average rainfall volume of 75.20 mm (average of two storms), while the remaining values range from 3.15 to 34.59 mm. At station 072a, there are two average rainfall volumes of 69.36 and 70.76 mm, while the rest range from 2.14 to 31.01 mm. The plot for station 020b (Figure 3.6.2) shows an unusual pattern. The strange behaviour of station 020b is not surprising since as mentioned previously, the monthly averages of rainfall volumes vary greatly from year to year and from month to month at this station. The plots of other stations do not show any obvious pattern for $\widetilde{pH}$ and $\tilde{V}$; there are a few outliers. Overall, the plots do not show any obvious relationship between the adjusted rainfall volumes and pH levels. Since we did not see any relationship in the adjusted data, we are convinced that there is no relationship between the rainfall volumes and the pH levels. It appears that the rainfall volume is not an important factor in explaining the pH levels.

In the light of the above preliminary examination, we conclude that Year, Station, and Month are three important factors in explaining the pH levels. The rainfall volumes do not appear to influence the pH levels. In the next section, using the individual (event) data, we examine how these factors influence the pH levels in detail. To investigate this issue, we first fit an ANOVA model to the event data to see if we can fit this data with a simpler model (i.e. fewer parameters).

3.3) An ANOVA Model

An ANOVA model would represent the pH values corresponding to the individual events as follows:

$$(1)\ pH_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + \epsilon_{ijkl},$$

where

pH_{ijkl} denotes the pH level of the l^{th} storm occuring at the i^{th} Year, j^{th} Station and k^{th} Month,

μ denotes the overall effect,

α_i denotes the Year effect, i = 1,..,6,

β_j denotes the Station effect, j = 1,..,9,

γ_k denotes the Month effect, k = 1,..,12,

$(\alpha\beta)_{ij}, (\alpha\gamma)_{ik}, (\beta\gamma)_{jk}$ denote the second order interactions,

$(\alpha\beta\gamma)_{ijk}$ denotes the third order interaction,

ϵ_{ijkl} denotes the residual.

This is a saturated ANOVA model in which the pH level is represented as a combination of two parts. The stochastic part is represented by ϵ and the remaining components of the model represent the deterministic part (where all factors are considered fixed). In the following sections, we estimate these two parts and then examine each estimate carefully. This examination is intended to reveal the stochastic structure of the data; for example, whether the residual variation can be reasonably approximated by the normal distribution. We would also like to capture the deterministic component of the pH level with a simpler model (i.e. one with fewer parameters).

3.3.1) Fitting the ANOVA model

As mentioned in chapter 1, the standard deviations of the pH levels are quite different from station to station; they also vary somewhat from year to year (Table 1.2). Therefore, it seems reasonable to allow the pH levels to have different variances for different year×station combinations; we assume that all pH levels measured in the same year and same station have the same variance. Using the data in each year and each station (about 50 to 100 data points), we can get an estimate of the variance (as presented in Table 1.2). However, since we know that monthly effects are important, we can get a better estimate by utilizing this knowledge about the structure of the data. Given a fixed year and a specific station, we can write the pH level of the l^{th} storm observed in the k^{th} month as:

$$pH_{kl} = \mu_k + \epsilon_{kl},$$

where

μ_k denotes the monthly effect, k = 1,..,12,

ϵ_{kl} denotes the residual.

Since we assume that the residuals in each year×station combination, have the same variance, σ^2, we get an unbiased estimator S^2 for σ^2, given by

$$S^2 = \frac{\sum_k \sum_l (pH_{kl} - p\bar{H}_{k.})^2}{\sum_k (n_k - 1)} \tag{2}$$

where

$p\bar{H}_{k.}$ denotes the average of all pH levels in the k^{th} month,

n_k denotes the number of observations in the k^{th} month.

These estimates are a bit smaller than those adjusted only for the yearly average instead of the monthly average. We present both estimates of the variances in Table 2.2 where the estimates adjusted for monthly average are denoted by sd(adj).

With these estimates of the variances for the pH levels, we use a weighted least squares (WLS) method to fit the model (1). There are two ways of fitting this model. The first is to consider the estimates of the variances obtained above as the true values; the second is to carry out re-weighted least squares in which the variances are re-estimated after each iteration until they do not change substantially. Here, to remain consistent with our assumptions about the variances, we use the former and think that the above estimates are already very close to the true values. Since the primary objective of this fitting is to obtain the estimated values for the deterministic components, not to estimate the variances, the use of the re-weighted least squares method seems unnecessary.

In what follows, the saturated model and various submodels will be fitted. To convince ourselves that the estimated values of the parameters would not be much different with re-weighting, we carried out the re-weighted least squares with one simple submodel consisting of Year, Month, and Station effects only:

$$pH_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + \epsilon_{ijkl}.$$

Our previous estimates of the variances were used as initial values and the maximum likelihood method was used to re-estimate the parameters and the variances, assuming normal data. The fitting procedure converged after three iterations (the absolute change in the estimated values of each parameter is less than .01) and the maximum of the differences between our estimated values of parameters (i.e. iteration 1) and the resulting values from the re-weighting method is less than .02 for each parameter. This supports our belief that re-weighted least squares is unnecessary.

Using the procedure GLM in SAS with weights w_{ij} defined as

$$w_{ij} = \frac{1}{S_{ij}^2},$$

where S_{ij}^2 is the estimated variance for year i at station j, calculated using (2), we get the following results (Y=Year, M=Month, S=Station):

Source	df	SS	MS
Y	5	81.6	16.32
M	11	366.7	33.34
S	8	222.4	27.80
YM	55	311.6	5.77
YS	29	198.3	6.84
MS	88	172.0	1.96

YMS	319	521.8	1.64
Error	2400	2449.1	1.02
Total	2915	4323.5	

Note that by using the WLS method to fit the parameters, we assume that pH_{ijkl}/S_{ij} has a variance of 1. Our estimated value of the variance (MS_E) is 1.02, which is close to 1. This indicates that our estimates for the variances used in the fitting are appropriate.

In this table, df denotes the degrees of freedom of the effect and SS is the reduction in the weighted residual sum of squares due to the source (factor), obtained by adding these sources sequentially to the model in the order listed. More precisely, let RSS(f) be the weighted sum of squares of residuals when fitting the model denoted by f:

$$RSS(f) = \sum_{ijkl} w_{ij}(pH_{ijkl} - \hat{f})^2,$$

where $\hat{f}$ denotes the WLS fitted model.

Then each SS can be written in terms of these RSS's. For example, the sum of squares due to the factor, Month, SS_M, is given by

$$SS_M = RSS(\mu + \alpha_i + \beta_j) - RSS(\mu + \alpha_i + \beta_j + \gamma_k).$$

The corresponding MS is just the ratio of SS and df. The sums of squares due to the sources in the above table are not substantially influenced by the order in which the components are sequentially added to the model. Specifically, provided that all 1^{st} order effects are entered, essentially the same SS's for higher order effects are obtained using different orders. The results in the table correspond to what seems the most natural order. Using the above results, we examine the deterministic and stochastic components in the next sections.

3.3.2) Deterministic component

In the saturated model (1), we use 516 parameters to capture the deterministic component. In the hope of finding a simpler model (one with fewer parameters) which adequately represents the deterministic components, we examine whether we can drop some effects from the full model. To examine this question, we compare different submodels containing combinations of the effects and interactions.

It is obvious that the main effects are very important (the mean square for each of the main effects is at least 16.6, while the largest mean square for the second order interactions is at most 6.8). Moreover, it is reasonable to include a higher order interaction term in the model only when the lower term is also included. For example, if the year×month interaction is in the model then the year and month effects must be in the model. Guided by these criteria, we consider only eight specific submodels and the full model. We use the log-linear convention to identify the submodels; for example, [Y,SM] represents the model containing Y, S, M, and SM. The results comparing the models to be considered are provided in the following table (p = number of parameters):

Submodel	p	RSS	R^2	$\tilde{R}^2$	$C_p - p$	B_1	A_1
[Y,M,S]	25	3652.8	.16	.15	761.8	3852.3	3702.8
[YM,S]	80	3341.2	.23	.21	505.2	3979.4	3501.2
[YS,M]	54	3454.5	.20	.19	592.5	3885.3	3562.5
[Y,MS]	113	3480.8	.20	.16	677.8	4382.3	3706.8
[YM,YS]	109	3142.9	.27	.25	335.9	4012.5	3360.9
[YM,MS]	168	3169.2	.27	.22	421.2	4509.5	3505.2
[YS,MS]	142	3282.5	.25	.20	508.5	4415.4	3566.5
[YM,YS,MS]	197	2970.9	.31	.26	251.9	4542.6	3364.9
[YMS]	516	2449.1	.43	.31	49.1	6565.7	3481.1

In this table, the values of various statistics often used as criteria for choosing the 'best' submodel are also included. These criteria, which are similar to those suggested in Weisberg (1980) for independent and identically distributed (iid) residuals, are described in what follows.

a) Multiple correlation coefficient

The square of the multiple correlation coefficient R^2 is often used as a criterion for comparing models. In this case, since we have only independent but not identically distributed residual data, we use an analogue of R^2, which is calculated by using the weighted sums of squares, for comparing submodels. A computing formula for R^2 in a p-parameter model (denoted by R_p^2) is:

$$R_p^2 = 1 - \frac{RSS_p}{SYY},$$

where

RSS_p is the weighted residual sum of squares using the submodel with p parameters,

SYY is the total sum of squares, given by

$SYY = \sum_{ijkl} w_{ij}(pH_{ijkl} - p\bar{H}_{....})^2$,

where

$p\bar{H}_{....}$ denotes the overall average,

w_{ij} is as previously defined.

Note that the large values of R_p^2 indicate that more variability is explained by the model. A disadvantage of this method is that it is only useful for comparing different models with a fixed number of parameters.

b) Adjusted R^2

To adjust the method in (a) for the stated disadvantage, corresponding to suggestions in Weisberg (1980), one may use an adjusted version of R^2, denoted by $\tilde{R}^2$ and defined by

$$\tilde{R}_p^2 = 1 - \frac{(n-1)}{(n-p)}(1 - R_p^2),$$

where n is the total number of observations.

Note that as before large values of $\tilde{R}^2$ indicate that more variability is explained by the model. The adjusted value can be negative.

c) Mallows' C_p statistic

Let $\widehat{pH}_{ijkl}$ be the predicted value of the corresponding observation obtained by the submodel. Let the mean square error (mse) be defined as

$$mse(\widehat{pH}_{ijkl}) = var(\widehat{pH}_{ijkl}) + [E(\widehat{pH}_{ijkl}) - E(pH_{ijkl})]^2.$$

The submodel which makes the weighted mean square errors, given by $w_{ij} \times mse(\widehat{pH}_{ijkl})$, as small as possible is preferable. Note that Weisberg (1980) suggests this criterion only for the iid case. However, since we use the weighted mean square error criterion, we in fact have an iid case where pH_{ijkl}/S_{ij} is the raw data with variance equal to

1. The Mallows' C_p statistic for a p-parameter model is defined by

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

where RSS_p is as previously defined. Mallows (1973) suggests that good models will have negative, or small, values of $C_p - p$.

We calculate R^2, $\tilde{R}^2$, C_p with SYY=4323.5 and n=2916, and present the results in the above table. All three methods suggest that the saturated model is the best one among those submodels considered. This suggests that all terms in the full model are needed to account for the pH levels. Note that the full model corresponds to using the monthly average as the fitted value for the pH level of any particular event.

Our objective in investigating the deterministic component is to reduce the full model to a model with fewer parameters which still captures the structure in the deterministic component of the data. To find a simpler model, we would first have to examine whether any of the different sources (i.e. Y, S, M, YS,...) could be dropped from the full model without serious effect on its ability to capture the deterministic component. This would hopefully lead us to a submodel. Subsequently, we would investigate the possiblity of replacing the remaining parameters with fewer parameters by using simple (possibly nonlinear) functional forms to represent some of the effects which were identified as important (a periodic form to represent any seasonality evident in the estimated Y, M, and YM effects, for example).

Using the MS due to the sources, we can see if any source can be eliminated. As noted earlier, the MS due to the main effects are very important; they range from 16.32 to 33.34. With $MS_{error} \simeq 1.02$, rough F tests would yield very small p-values for these main effects. The fate of interaction terms are not so clear. The MS due to the year×month (MS_{YM}) and year×station (MS_{YS}) interactions are considerably smaller than the MS due to the main effects (the values are 5.77 and 6.84), but rough F tests would yield small p-values. The MS due to the month×station (MS_{MS}) and year×station×month (MS_{YSM}) interactions are much smaller than MS_{YM} and MS_{YS} ($MS_{MS} = 1.96$, $MS_{YSM} = 1.64$); intuition suggests that these interactions are not really important. Although the rough F tests would still yield statistical significance for these interactions, the significance in this case may be inevitable and due to the very large number of observations available (2916). The results using different criteria (described above) also suggest that every term in the model is important in explaining the pH levels (i.e. the full model is the 'best' one). Note that none of these methods 'adjust' for the large number of observations available even though the latter plays an important role in these results.

To pursue this issue further, we can use the criteria proposed by Schwarz (1978) and Akaike (1973) for comparing the fit of different models. Let p be the number of estimable parameters in a model. Schwarz proposed that the model which maximizes

$$\mathrm{B} = \log(\text{maximum likelihood}) - \tfrac{p}{2} \log(\mathrm{n})$$

be chosen. Since the number of observations (n) is fixed, with the assumption of independent and normal data having known variances (i.e. $\sigma_{ij}^2 = S_{ij}^2$), the criterion is equivalent to minimizing

$$\mathrm{B}_1 = RSS_p + p \times \log(\mathrm{n}).$$

Akaike proposed that the model which maximizes

$$\mathrm{A} = \log(\text{maximum likelihood}) - p$$

be chosen. With the same assumptions, this is equivalent to minimizing

$$\mathrm{A}_1 = RSS_p + 2 \times p.$$

Note that log(n) = 7.978, so $B_1 = A_1 + 5.978 \times p$ in this case.

The values of A_1 and B_1 for the submodels are presented in the above table. According to Akaike's criterion, the submodel [YS,YM] is the best; the last two terms in the full model (i.e. SM, YSM) can be dropped. This supports our earlier suggestion that MS_{SM} and MS_{YSM} are not really important. Schwarz's criterion, however, prefers the submodel [Y,S,M] to the rest. Nishii (1984) proves that the criteria proposed by Mallows and Akaike are asymptotically equivalent under general conditions; asymptotically, they have a positive probability of selection only for models that properly include the true model. However, the Schwarz criterion behaves somewhat differently; the model chosen by the Schwarz criterion is a 'consistent' estimator of the true model. To a certain extent, this behaviour is evident in our analysis. The three 'best' submodels chosen by the Mallows and Akaike criteria are the same (although their rankings are slightly different), but the Schwarz criterion indicates that smaller submodels are preferable. Since the term log(n) adjusts for the large number of observations available, the Schwarz criterion may be more appropriate in this present context than those of Akaike and Mallows. These results clearly suggest that the deterministic component of the pH levels can be captured by some simpler model than the saturated one.

The model [YS,YM] chosen by Akaike's criterion is already much simpler than the full model (109 versus 516 parameters). However, according to the criterion proposed by Schwarz, an even simpler model can possibly be used. Nishii's results suggest that although the submodel chosen by the Schwarz criterion is an asymptotically 'consistent' estimator of the true model, the submodel chosen by the Akaike criterion has a greater likelihood of properly containing the true model. Therefore, to find the smallest possible model for the pH levels, we would need to examine the estimated values of the parameters in the model [YS,YM] to see whether they can be approximated by a smoothed function. This smoothed function (if found) would be a proposed model and its properties and fit to the data could be further examined. However, due to limitations of time, this interesting project must be deferred.

3.3.3) Stochastic component

Many methods of analysis employed in environmetrics are based on assumptions of normality. Therefore, it is important to check the normality of the pH levels. The simplest way of examining this question is to use the normal probability plot. The residuals from the full model of the previous section are standardized and plotted for each station (all years together) separately. Using SAS, we calculate the normal probability by

$$Pr(\epsilon_i) = \Phi^{-1}(r_i - \frac{3}{8})/(n + \frac{1}{4}),$$

where

ϵ_i denotes the residual of the i^{th} observation,

r_i denotes the rank of ϵ_i,

n denotes the total number of residuals,

Φ denotes the cumulative normal distribution.

The resulting plots for the different stations are presented in Figures 3.7.1 to 3.7.9. The plots of stations 013a, 044a, 171b are presented in 3.7.1, 3.7.4, 3.7.9 respectively. These plots do not suggest any violation of the normality assumption.

Figure 3.7.3 represents the normal plot of data from station 043a. The plot is very well-behaved, except for one possible outlier: pH=5.95 in November 1980 (the rest of the pH readings in that month range from 4.02 to 4.35). The normal plot for station 048a (Figure 3.7.5) has a similar property. It also has one possible outlier: pH=6.00 in December 1981 (the rest of the pH readings in that month range from 3.70 to 4.71). The plot of station 072a (Figure 3.7.8) does not suggest any serious violation of the normality assumption; however, there is one apparent outlier: pH=6.57 in June 1981 (the rest of the pH readings in that month range from 3.86 to 4.07).

Figures 3.7.6 and 3.7.7 represent data from stations 057a and 065a respectively. The plots do not show any serious violation of the normality assumption. However, these plots have some disturbing features. Although most stations have collinear data, there are some points on both tails which are slightly off those straight lines. A more careful examination of the data from these stations is described below.

Figure 3.7.2 represents the normal plot of data from station 020b. The normal plot shows a serious violation of the normal assumption. Instead of having a roughly linear, a curve is obtained; the data needs to be examined more carefully.

We can investigate these 'questionable' stations further by examining the normal plots for each year in each of these stations separately. The plots (not included in this work) for stations 057a and 065a for each year separately show that except for some outliers in the data, there are no obvious violations of the normality assumption. However, the plots of station 020b suggest a serious departure from normality. The plots for 1978 to 1982 are presented in Figures 3.8.1 to 3.8.5 respectively. Except for Figure 3.8.1, the plots clearly exhibit the extreme skewness of the data from this station. Figure 3.8.1 shows that the data for 1978 support the normality assumption except for some apparent outliers.

These results indicate that the stochastic components of the pH levels obtained at station 020b cannot be considered as arising from a normal distribution. However, except for a few possible outliers, the data from the remaining stations can be reasonably approximated as normal.

3.4) Conclusion

In this chapter, we have examined the data set carefully in an attempt to uncover underlying patterns which might be used to find an appropriate model for the pH levels. The results indicate that there is a seasonal structure in the pH levels. Specifically, the time effects decrease slowly from January to August, increase sharply from September to November, and then decrease a little bit in December. The rainfall volumes, however, do not show any obvious seasonal structures. There is no obvious effect of the rainfall volumes on the pH levels. The station effects are quite important in explaining the pH levels.

The deterministic components of the pH levels were examined by fitting a saturated ANOVA model; the importance of the individual components in the model were also investigated. The results suggest that the deterministic components can be captured by a smaller model than the full model.

The stochastic component was examined for the appropriateness normality assumptions by using normal probability plots. The results indicate that serious violations of the normality assumption are found in the data obtained at station 020b. However, except for a few possible outliers, the data from other stations can be reasonably approximated as normal.

4) CONCLUSION

In this study, we have learned that the model developed by Eynon and Switzer (1983) for analyzing pH levels, does not seem to be appropriate for the data obtained from the MAP3S/PCN montoring network. However, the general approach seems to be useful in analyzing the spatial-temporal data. For this reason, the approach including the interpolation step called kriging was completely demonstrated using the Eynon-Switzer model.

In an effort to identify a more appropriate model for the data, we examined the raw data in detail. A full ANOVA model, including the three factors, Year, Month, and Station, was fitted to the individual pH levels. The results suggested that the residuals have different variances at different stations. The normality of the residuals was examined. The conclusion is that the data from all stations except one can reasonably be approximated as coming from normal distributions. This is a very useful result because many methods of analysis employed in environmetrics are based on assumptions of normality.

Different criteria such as Akaike, Schwarz, Mallows, etc, are used to identify the 'best' submodel (i.e. fewer parameters than the full model) for capturing the deterministic component of the data. The results suggested that it is possible to capture the deterministic component by a much smaller model than the full model.

It would be very interesting to examine whether the 'best' submodel can be represented by a smoothed function. The abilities to fit the data and the properties of this smoothed function (if found) could be further examined. Moreover, since the variances of the residuals of the data are not homogeneous from station to station, we must have the estimates of the variances of the residuals at any locations in order to do interpolation (kriging in this case). The question of whether we would be able to estimate these variances is also interesting. However, due to limitations of time, these interesting projects must be deferred.

REFERENCES

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *2^{nd} International Symposium on Information Theory.* (B. N. Petrov and F. Czaki, eds.). Akademiai Kiado, Budapest.

Delhomme, J. P. (1978). Kriging in the hydrosciences. *Adv. Water Resources*, **1**, 251-266.

Egbert, G. D. and Lettenmaier, D. P. (1986). Stochastic Modeling of the Space-Time Structure of Atmospheric Chemical Deposition. *Water Resources Research*, **22**, 165-179.

Eynon, B. P. and Switzer, P. (1983). The variability of rainfall acidity. *The Canadian Journal of Statistics*, **11**, 11-24.

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

Journel, A. G. and Huijbregts, Ch. J. (1978). Mining Geostatistics. London: Academic Press.

Nishii, R. (1984). Asymptotic properties of criteria for selection of variable in multiple regression. *Annals of Statistics*, **12**, 758-765.

Schwarz, G. (1978). Estimating the dimension of model. *Annals of Statistics*, **6**, 461-464.

Weisberg, S. (1980). Applied Linear Regression. New York: Wiley.

Table 1.1:

Site location : The MAP3S/PCN monitoring network.

ADS site identity	Location	Latitude	Longitude	Elevation (meters)	First active date	Years of data
013a	Lewes, Delaware	38 46 00	75 00 00	0	01-Mar-78	4
020b	Illinois, Illinois	40 03 12	88 22 19	212	20-Nov-7	5
043a	Whiteface, New York	44 23 26	73 51 34	610	11-Oct-76	6
044a	Ithaca, New York	42 24 03	76 39 12	509	26-Oct-76	6
048a	Brookhaven, New York	40 52 00	72 53 00	25	09-Feb-78	5
057a	Oxford, Ohio	39 31 51	84 43 25	284	01-Oct-78	4
065a	Penn State, Pennsylvania	40 47 18	77 56 47	393	22-Sept-76	6
072a	Virginia, Virginia	38 02 23	78 32 31	172	12-Dec-76	5
171b	Oakridge, Tennessee	35 57 41	84 17 14	341	07-Jan-81	2

* A. Olsen and C. Watson, "Acid Deposition System (ADS) for Statistical Reporting", EPA-600/8-84-023, September 1984, pp. C.7-C.13.

Table 1.2: Summary of data.

ADS Identity	Year	No. of rainfalls	pH			Volumes (mm)	
			Mean	sd	sd(adj)	Mean	sd
013a	79	52	4.37	.40	.28	24.99	29.49
	80	71	4.25	.35	.30	12.88	13.92
	81	64	4.22	.40	.36	13.16	12.97
	82	73	4.26	.55	.48	13.31	15.28
020b	78	50	4.25	.55	.58	10.44	12.92
	79	40	4.73	.89	.77	19.80	29.48
	80	94	4.28	.68	.69	9.21	11.65
	81	87	4.34	.65	.57	10.68	14.26
	82	80	4.37	.52	.48	11.12	14.01
043a	77	83	4.24	.33	.32	12.97	13.18
	78	54	4.16	.31	.24	14.16	12.39
	79	50	4.08	.18	.17	17.08	14.39
	80	87	4.03	.29	.26	9.52	7.78
	81	99	4.06	.23	.23	10.92	16.10
	82	98	4.21	.39	.37	8.22	10.06
044a	77	52	4.16	.28	.20	19.06	14.20
	78	60	4.05	.27	.20	13.83	11.09
	79	62	4.08	.29	.27	17.80	19.63
	80	75	4.19	.32	.27	12.41	12.20
	81	72	4.10	.23	.20	16.49	17.06
	82	80	4.14	.33	.31	11.54	12.01
048a	78	48	4.07	.35	.30	14.86	15.21
	79	71	4.10	.49	.42	12.93	18.15
	80	64	4.14	.51	.44	12.61	12.77
	81	80	4.18	.45	.39	11.93	13.26
	82	69	4.23	.45	.40	15.49	18.55
057a	79	47	4.22	.23	.20	15.08	16.97
	80	75	4.08	.29	.24	11.72	13.34
	81	75	3.80	.24	.19	9.67	8.42
	82	83	3.99	.37	.30	11.20	13.20
065a	77	80	4.06	.27	.22	14.93	17.84
	78	70	4.04	.30	.25	13.00	11.03
	79	61	4.20	.27	.24	18.57	18.57
	80	40	4.14	.61	.20	12.46	12.61
	81	69	4.19	.30	.25	11.64	15.24
	82	86	4.23	.36	.33	9.46	10.85
072a	78	57	4.06	.34	.28	18.00	17.47
	79	51	4.17	.30	.29	25.30	32.42
	80	65	4.09	.30	.31	10.85	11.50
	81	56	4.15	.46	.48	14.98	21.79
	82	58	4.15	.30	.29	16.01	14.74
171b	81	65	4.16	.23	.23	14.63	13.75
	82	64	4.27	.23	.22	23.14	20.17

TABLE 2.1:

The fitted constant coefficients obtained by fitting the Eynon-Switzer model

Year	Para-				Station					
	meter	013a	020b	043a	044a	048a	057a	065a	072a	171b
	a			0.0185	-0.0015			0.0598		
77	b			0.1375	0.1855			0.2364		
	c			0.0466	0.0422			0.0348		
	d			4.1311	3.9918			3.9618		
	a		-0.1100	0.2063	0.0218	-0.0477		0.0474	0.0941	
78	b		-0.0066	0.1897	0.1631	-0.0507		0.1652	0.0257	
	c		-0.0064	0.0227	0.0763	0.0662		0.1069	0.0477	
	d		4.2678	4.1145	3.8588	3.8740		3.8083	3.9056	
	a	-0.1478	-0.2415	-0.0241	-0.1017	0.0742	-0.0466	-0.0045	-0.0871	
79	b	0.2606	0.3583	0.0729	0.1117	0.3195	-0.0081	0.0871	0.0807	
	c	0.0187	-0.0289	0.0348	0.0565	0.0337	-0.0056	0.0445	0.0254	
	d	4.2778	5.0538	3.9550	3.8905	4.0563	4.2408	4.0607	4.0381	
	a	-0.0616	-0.1810	-0.0446	0.0366	-0.1499	0.2058	0.0586	-0.0176	
80	b	0.1467	0.3039	0.1561	0.1855	0.1075	0.0729	0.3813	0.1559	
	c	0.0906	-0.0015	0.0285	0.0449	0.1371	-0.0020	0.1805	0.0665	
	d	4.0414	4.3846	3.9929	4.0731	3.8802	4.0784	3.9236	3.9623	
	a	0.1242	-0.0530	0.0252	0.0412	0.1368	0.1531	-0.0945	0.0771	0.0771
81	b	0.2258	0.3821	0.0365	0.1830	0.1232	0.0367	0.1496	-0.0399	0.0510
	c	0.0450	-0.0004	0.0553	0.0311	0.1431	0.0331	0.0447	0.0201	0.0495
	d	4.1085	4.3990	3.9527	4.0137	3.9080	3.7293	4.1141	4.0722	4.0176
	a	0.1082	-0.0193	-0.1840	-0.0383	-0.1048	-0.0392	0.1375	0.0399	0.0546
82	b	0.2331	0.0575	0.0494	0.2062	0.2176	0.1322	0.1694	0.1471	0.0882
	c	0.0855	0.0161	0.0690	0.1026	0.1375	0.0266	0.0516	0.0259	0.0087
	d	4.6606	4.3374	4.0859	3.9298	3.9194	3.9313	4.1205	4.0602	4.2359

Table 2.2; ANOVA and median decompositions on $\hat{a}, \hat{b}, \hat{c}$, and $\hat{d}$.

(1) The residuals and effects obtained by decomposing the estimated coefficients a(x) using AVERAGE (part a) and MEDIAN (part b).

(a) This uses AVERAGE for decomposition.

Ads id.	Year	013a	020b	043a	044a	048a	057a	065a	072a	171b	Effect
Resid.	77			-0.01	-0.02			0.00			0.02
	78		0.04	0.16	-0.02	-0.08		-0.04	0.02		0.05
	79	-0.08	-0.07	0.05	-0.02	0.17	-0.04	0.04	-0.03		-0.07
	80	-0.05	-0.06	-0.02	0.06	-0.11	0.16	0.05	-0.02		-0.02
	81	0.07	-0.00	-0.02	-0.00	0.11	0.04	-0.18	0.01	-0.05	0.05
	82	0.11	0.09	-0.18	-0.02	-0.08	-0.10	0.11	0.03	0.01	-0.01
											Overall average
Col. eff.		0.00	-0.10	-0.00	-0.01	-0.02	0.07	0.03	0.02	0.05	0.00

(b) This uses MEDIAN for decomposition.

Ads id.	Year	013a	020b	043a	044a	048a	057a	065a	072a	171b	Effect
Res.	77			0.01	-0.00			-0.01			0.02
	78		0.06	0.18	0.00	0		-0.04	0.03		0.03
	79	-0.10	-0.07	0.05	-0.02	0.22	-0.04	0.01	-0.05		-0.07
	80	-0.05	-0.05	-0.01	0.08	-0.04	0.17	0.04	-0.01		-0.03
	81	0.05	0	-0.02	0.01	0.17	0.04	-0.20	0	-0.04	0.05
	82	0.11	0.10	-0.16	-0.00	-0.00	-0.08	0.10	0.03	0.04	-0.02
											Overall median
Col. eff.		0.02	-0.10	-0.01	-0.02	-0.08	0.06	0.05	0.03	0.04	-0.00

(2) The residuals and effects obtained by decomposing the estimated coefficients b(x) using AVERAGE (part a) and MEDIAN (part b).

(a) This uses AVERAGE for decomposition.

Ads id.		013a	020b	043a	044a	048a	057a	065a	072a	171b	Effect
	Year										
Resid.	77			-0.01	-0.03			-0.00			0.04
	78		-0.16	0.15	0.06	-0.13		0.03	0.02		-0.07
	79	0.03	0.12	-0.05	-0.08	0.16	-0.08	-0.13	-0.00		0.02
	80	-0.11	0.04	0.01	-0.03	-0.08	-0.03	0.14	0.04		0.04
	81	0.03	0.18	-0.05	0.03	-0.00	-0.00	-0.03	-0.10	0.00	-0.02
	82	0.02	-0.16	-0.06	0.04	0.08	0.08	-0.03	0.07	0.02	-0.00
											Overall average
Col. eff.		0.07	0.07	-0.04	0.03	-0.00	-0.09	0.05	-0.07	-0.08	0.15

(b) This uses MEDIAN for decomposition.

Ads id.		013a	020b	043a	044a	048a	057a	065a	072a	171b	Effect
	Year										
Res.	77			-0.01	-0.04			0.01			0.05
	78		-0.25	0.13	0.02	-0.15		0.03	-0.06		-0.04
	79	0.07	0.11	0.01	-0.03	0.22	-0.02	-0.05	0		-0.03
	80	-0.10	0	0.04	-0.01	-0.05	0.00	0.19	0.01		0.02
	81	0.01	0.11	-0.05	0.02	0	-0.00	-0.01	-0.15	-0.01	-0.01
	82	-0.01	-0.24	-0.06	0.01	0.07	0.07	-0.02	0.01	0.01	0.01
											Overall median
Col. eff.		0.09	0.15	-0.04	0.04	0	-0.08	0.04	-0.01	-0.07	0.14

(3) The residuals and effects obtained by decomposing the estimated coefficients c(x) using AVERAGE (part a) and MEDIAN (part b).

(a) This uses AVERAGE for decomposition.

Ads id.		013a	020b	043a	044a	048a	057a	065a	072a	171b	Effect
	Year										
Resid.	77			0.01	-0.01			-0.03			-0.01
	78		-0.01	-0.02	0.01	-0.04		0.03	0.01		0.00
	79	-0.01	0.00	0.02	0.02	-0.04	0.01	-0.01	0.01		-0.03
	80	0.01	-0.02	-0.03	-0.03	0.01	-0.03	0.08	0.01		0.02
	81	-0.01	0.01	0.01	-0.03	0.04	0.02	-0.03	-0.01	0.02	-0.00
	82	0.02	0.01	0.02	0.03	0.02	0.00	-0.03	-0.02	-0.03	0.01
											Overall average
Col. eff.		0.01	-0.05	-0.01	0.01	0.05	-0.04	0.03	-0.01	-0.02	0.05

(b) This uses MEDIAN for decomposition.

Ads id.		013a	020b	043a	044a	048a	057a	065a	072a	171b	Effect
	Year										
Res.	77			-0.00	-0.01			-0.01			-0.01
	78		-0.01	-0.04	0.01	-0.07		0.04	0		0.01
	79	-0.03	-0.00	0.00	0.02	-0.08	-0.01	0.01	0.01		-0.02
	80	0.02	0	-0.03	-0.02	0	-0.03	0.12	0.02		0.01
	81	-0.02	0.01	0.01	-0.02	0.02	0.02	-0.01	-0.02	0.02	-0.00
	82	0.02	0.02	0.01	0.04	0.00	0.01	-0.01	-0.02	-0.02	0.00
											Overall median
Col. eff.		0.02	-0.06	0.00	0.01	0.08	-0.03	0.01	-0.01	-0.02	0.05

(4) The residuals and effects obtained by decomposing the estimated coefficients d(x) using AVERAGE (part a) and MEDIAN (part b).

(a) This uses AVERAGE for decomposition.

Ads id.		013a	020b	043a	044a	048a	057a	065a	072a	171b	Effect
	Year										
Resid.	77			0.13	0.07			0.00			-0.04
	78		-0.13	0.17	-0.01	0.04		-0.10	-0.01		-0.09
	79	0.02	0.43	-0.21	-0.20	-0.00	0.11	-0.07	-0.10		0.13
	80	-0.06	-0.08	-0.02	0.14	-0.02	0.11	-0.05	-0.02		-0.02
	81	0.02	-0.06	-0.06	0.08	0.01	-0.24	0.15	0.10	-0.08	-0.03
	82	-0.07	-0.16	0.04	-0.04	-0.02	-0.07	0.11	0.04	0.10	0.01
											Overall average
Col. eff.		0.06	0.42	-0.03	-0.11	-0.14	-0.07	-0.07	-0.06	0.06	4.07

(b) This uses MEDIAN for decomposition.

Ads id.		013a	020b	043a	044a	048a	057a	065a	072a	171b	Effect
	Year										
Res.	77			0.14	0.02			-0.02			-0.03
	78		0	0.22	-0.01	0.09		-0.07	0.02		-0.13
	79	0.16	0.62	-0.10	-0.15	0.11	0.22	0.02	-0.02		0.04
	80	-0.03	-0.01	-0.02	0.07	-0.03	0.10	-0.08	-0.05		-0.00
	81	0.03	0.00	-0.07	0.01	0	-0.25	0.11	0.05	-0.09	0.00
	82	-0.06	-0.10	0.02	-0.11	-0.03	-0.10	0.07	0	0.09	0.04
											Overall median
Col. eff.		0.06	0.38	0.00	-0.02	-0.11	-0.03	-0.01	-0.00	0.09	4.02

Table 2.3: The ratio of residual SS's of the corrected and uncorrected series of pH levels.

	Station								
Year	013a	020b	043a	044a	048a	057a	065a	072a	171b
1977			.86	.63			.50		
1978		.99	.55	.52	.56		.52	.79	
1979	.72	.86	.73	.48	.68	.98	.65	.74	
1980	.70	.90	.82	.73	.75	.72	.57	.68	
1981	.75	.85	.73	.61	.98	.73	.74	.96	.67
1982	.76	.98	.77	.55	.53	.89	.76	.81	.86
Average	.73	.92	.74	.59	.70	.71	.62	.80	.77

Table 2.4: Sums of squares errors for different models.

Station	Year	$SS_{E(d)}$	$SS_{E(d,a,b)}$	RSS^{*}_{d}	$SS_{E(full)}$	RSS^{**}_{full}
043a	1977	8.97	8.23	09.0	7.71	06.7
	1978	5.11	2.95	73.2	2.80	05.4
	1979	1.67	1.59	05.0	1.22	30.3
	1980	7.27	6.12	18.8	5.95	02.9
072a	1982	5.28	4.59	15.0	4.28	07.2
065a	1978	4.25	3.84	10.7	3.30	16.4
044a	1981	3.82	2.86	33.6	2.34	22.2
013a	1980	8.54	7.80	08.5	5.71	36.6

$SS_{E(d)}$ denotes the sum of squares error corresponding to fitting a reduced model which involves only term d (i.e. pH = d).

Similarly, $SS_{E(d,a,b)}$ corresponds to the model

$$pH = a\sin\frac{2\pi t}{365} + b\cos\frac{2\pi t}{365} + d.$$

Note : the full model is

$$pH = a\sin\frac{2\pi t}{365} + b\cos\frac{2\pi t}{365} + log_{10}\frac{c\cdot I}{1-\exp(-c\cdot I)} + d.$$

* denotes ($SS_{E(d)}$ - $SS_{E(d,a,b)}$) / $SS_{E(d,a,b)}$ (in percent).

** denotes ($SS_{E(d,a,b)}$ - $SS_{E(full)}$) / $SS_{E(full)}$ (in percent).

Table 2.5:

The results of fitting the temporal variogram $V_\beta(w)$ to the data, where

$$V_\beta(w) = \frac{g_1}{1 + [w^2/g_2]^{-1}}.$$

Station	Year	Number of data pts.	$\hat{g}_1$	se($\hat{g}_1$)	$\hat{g}_2$	se($\hat{g}_2$)
013a	1979	1326	.116	.004	2.93	3.54
013a	1980	2085	.085	.002	2.04	2.37
072a	1982	1653	.079	.003	2.85	3.72
171b	1982	2016	.060	.002	1.54	2.60

Table 2.6: Temporal variogram estimated with data of each year and each station using

$$V_\beta = \hat{g}_1 = ave[\frac{1}{2}(p\tilde{H}(x_i,t_j) - p\tilde{H}(x_i,t_{j'}))^2],$$

where t_j and $t_{j'}$ belong to the same year, and i denotes station i.

	Station								
Year	013a	020b	043a	044a	048a	057a	065a	072a	171b
1977			.094	.050			.038		
1978		.326	.056	.040	.100		.053	.094	
1979	.115	.805	.026	.051	.168	.057	.051	.068	
1980	.085	.430	.071	.075	.208	.080	.242	.063	
1981	.122	.364	.042	.033	.153	.043	.066	.204	.036
1982	.234	.275	.117	.069	.125	.122	.098	.078	.060

Table 2.7 :·

The residuals and effects obtained by decomposing the estimated coefficients g1 using AVERAGE (part a) and MEDIAN (part b).

(a) This uses AVERAGE for decomposition.

Ads id. Year	013a	020b	043a	044a	048a	057a	065a	072a	171b	Effect
Resid. 77			0.10	0.07			0.02			-0.07
78		-0.09	0.01	0.01	-0.03		-0.02	0.01		-0.02
79	-0.06	0.33	-0.08	-0.04	-0.02	-0.05	-0.08	-0.07		0.04
80	-0.08	-0.03	-0.02	-0.00	0.03	-0.02	0.13	-0.06		0.02
81	-0.00	-0.06	-0.01	-0.01	0.02	-0.02	-0.01	0.12	0.00	-0.01
82	0.10	-0.16	0.05	0.02	-0.02	0.05	0.01	-0.02	0.01	-0.00
										Overall average
Col. eff.	0.01	0.31	-0.06	-0.08	0.02	-0.06	-0.04	-0.03	-0.08	0.13

(b) This uses MEDIAN for decomposition.

Ads id. Year	013a	020b	043a	044a	048a	057a	065a	072a	171b	Effect
Res. 77			0.06	0.02			-0.00			-0.02
78		-0.05	-0.00	-0.01	-0.06		-0.01	0.01		0.00
79	-0.00	0.45	-0.02	0.01	0.02	-0.00	0.00	0		-0.01
80	-0.05	0.05	0.00	0.01	0.04	0.00	0.17	-0.03		0.01
81	0.00	0	-0.01	-0.01	0	-0.02	0.01	0.13	0.01	-0.00
82	0.06	-0.14	0.02	-0.03	-0.08	0.01	-0.01	-0.05	-0.01	0.05
										Overall median
Col. eff.	0.05	0.30	-0.02	-0.02	0.09	-0.00	-0.01	0.01	-0.04	0.07

Table 2.8:

The estimated value of α in year i is

$$\hat{\alpha}(x_i) = \frac{\sum_j \widetilde{pH}(x_i,t_j)}{n_i}.$$

	Year			
Station	1979	1980	1981	1982
013a	4.28	4.11	4.14	4.14
020b	4.84	4.28	4.33	4.27
043a	4.01	3.94	4.00	4.12
044a	4.00	4.05	4.02	4.03
048a	4.09	4.03	4.13	4.10
057a	4.16	3.95	3.72	3.89
065a	4.14	4.13	4.14	4.13
072a	4.06	3.98	4.05	3.99
171b			4.07	4.07

Table 3.1:

Station effects, using ANOVA and Median decompositions (pH levels).

Station	Year											
	1977		1978		1979		1980		1981		1982	
	Anova	Median	Anova	Median	Anova	Median	Anova	Median	Anova	Median	Anova	Median
013a					.09	.22	.07	.06	.08	.06	.08	.15
020b			.11	.15	.64	.31	.24	.12	.25	.13	.16	.16
043a	.09	.10	.11	.14	-.19	-.17	-.12	-.12	-.07	-.11	-.01	.00
044a	-.01	.00	-.06	-.04	-.19	-.17	.00	.07	-.03	.07	-.06	-.06
048a			-.04	-.02	-.13	.03	.02	.06	.03	.00	.02	.03
057a					-.04	.02	-.09	-.07	-.30	-.33	-.22	-.25
065a	-.08	-.07	-.05	-.08	-.05	-.02	-.04	-.12	.03	.13	.01	-.02
072a			-.06	.02	-.09	-.09	-.08	-.06	-.01	-.05	-.06	-.10
171b									.02	.04	.06	.04

Table 3.2:

Monthly effects, using ANOVA and Median decompositions (pH levels).

	Year											
	1977		1978		1979		1980		1981		1982	
	Anova	Median	Anova	Median	Anova	Median	Anova	Median	Anova	Median	Anova	Median
January	.02	.09	.38	.40	.02	.18	.24	.20	-.09	-.11	.05	.06
February	-.13	-.22	-.10	-.19	-.07	.02	-.05	-.03	.18	.18	.01	.01
March	.26	.20	.11	.12	-.09	-.06	.12	.14	.14	.10	-.07	-.05
April	.07	.09	.01	-.01	-.02	-.02	.03	.01	-.03	-.01	.15	.09
May	-.22	-.25	.10	.15	-.04	-.06	-.18	-.04	-.05	-.07	-.16	-.18
June	-.11	-.12	-.11	-.17	-.18	-.11	-.14	-.11	-.08	-.13	-.01	-.02
July	-.31	-.34	-.07	-.11	-.29	-.22	-.06	-.12	-.09	-.13	-.18	-.25
August	-.13	-.10	-.08	-.02	-.22	-.12	-.32	-.24	-.17	-.17	-.21	-.26
September	-.03	-.06	-.12	-.10	.18	.26	-.06	-.08	-.05	-.09	-.04	-.14
October	.13	.11	-.01	.05	.11	.03	.28	.18	-.07	.07	.08	.01
November	.20	.23	-.08	.10	.41	.19	.12	.19	.15	.04	.17	.09
December	.28	.29	.11	.09	.18	.08	.00	.01	.02	.05	.22	.17

Table 3.3:

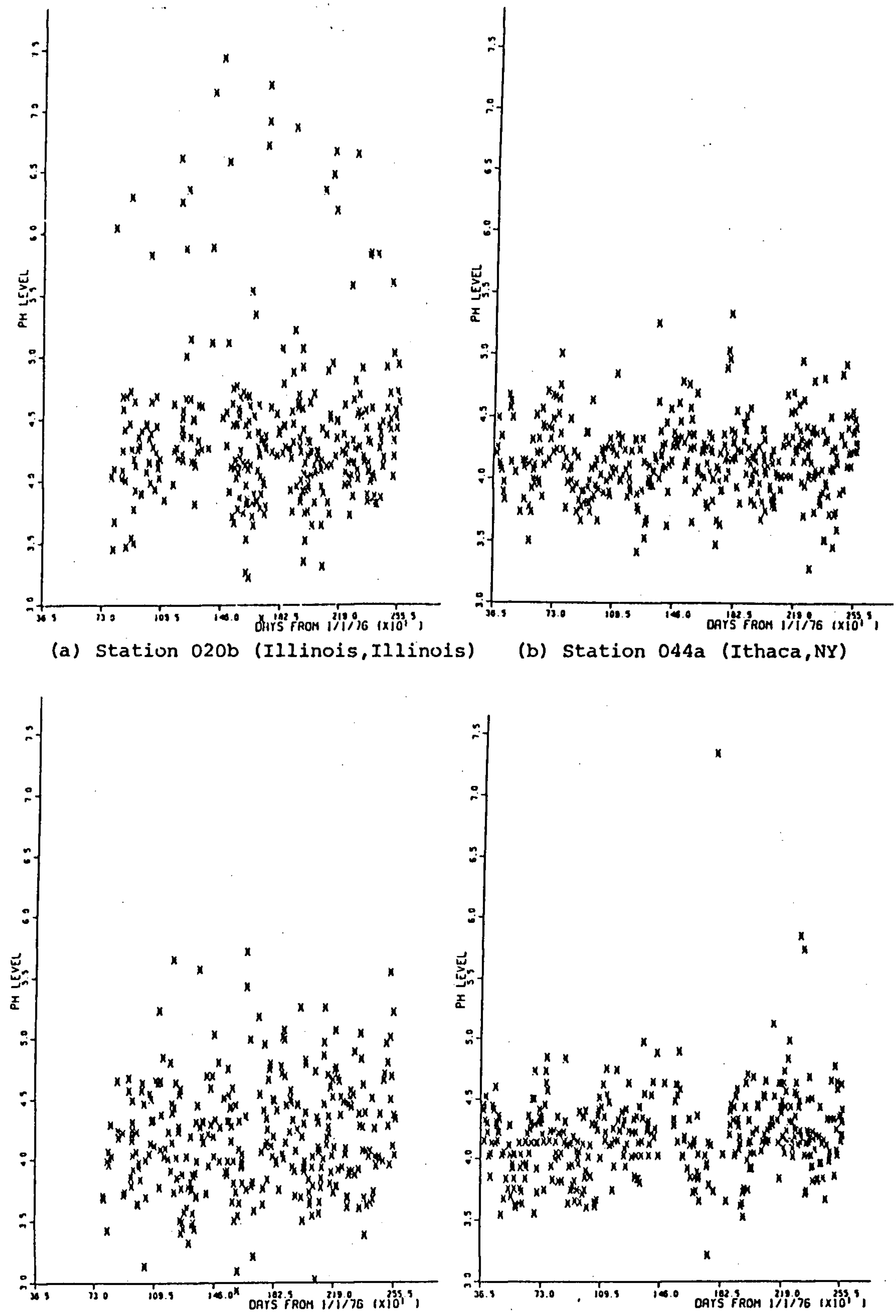
Station effects on rainfall volumes, using ANOVA and Median decompositions.

Station	Year											
	1977		1978		1979		1980		1981		1982	
	Anova	Median	Anova	Median	Anova	Median	Anova	Median	Anova	Median	Anova	Median
013a					11.39	11.68	2.29	5.09	0.31	1.78	0.57	3.09
020b			-3.88	-4.51	0.40	-9.25	-2.80	-3.07	-2.61	-2.26	-2.02	-0.11
043a	-2.46	-2.89	0.55	0.48	-2.48	0.38	-0.54	-0.36	-1.24	-1.96	-5.10	-4.31
044a	3.00	6.61	-0.51	0.37	-2.20	-0.05	1.29	1.12	4.81	6.31	-1.64	0.00
048a			1.05	-0.37	-8.01	-5.69	1.09	2.95	-1.09	0.00	1.99	3.93
057a					-3:44	-2.66	-0.27	-0.74	-2.94	-1.29	-2.28	-2.31
065a	-0.54	0.00	-1.28	-2.03	-2.09	0.05	-0.52	-0.43	-1.86	-1.97	-4.05	-2.40
072a			3.84	5.63	6.45	4.18	-0.58	0.37	2.31	1.42	2.17	1.89
171b									2.33	3.73	10.53	11.20

Table 3.4:

Monthly effects on rainfall volumes, using ANOVA and Median decompositions.

Month	Year											
	1977		1978		1979		1980		1981		1982	
	Anova	Median	Anova	Median	Anova	Median	Anova	Median	Anova	Median	Anova	Median
January	-9.21	-10.62	10.16	9.67	7.47	10.66	-1.83	-3.69	-8.95	-7.99	3.14	1.80
February	-9.08	-7.20	-2.63	-2.10	9.38	1.25	-3.70	-4.32	2.30	3.61	2.44	1.11
March	4.41	2.75	-3.47	-0.67	-1.93	1.94	3.57	3.34	-4.81	-4.26	-0.62	-1.70
April	2.74	1.72	-3.57	-1.86	-0.78	1.54	3.34	2.38	-1.08	-1.03	-1.66	-2.04
May	-10.09	-7.24	8.98	12.50	0.25	-0.24	3.97	4.67	-0.84	-0.54	-1.84	-2.70
June	2.46	-1.20	-4.60	-5.57	-1.21	0.46	0.31	0.06	-0.41	-1.66	1.41	1.64
July	-5.30	-2.63	-0.34	-0.98	-5.11	-3.68	-1.96	-2.31	4.77	10.10	1.38	1.86
August	-0.50	-3.11	0.57	1.43	3.24	-3.03	2.31	2.64	2.65	6.26	0.45	1.60
September	7.25	6.69	-1.71	-0.20	3.64	1.64	-2.09	-1.51	4.48	4.61	-1.63	-0.98
October	9.93	10.50	-2.30	1.69	-5.60	-0.45	2.72	1.77	5.94	5.30	-3.01	-3.41
November	4.04	3.25	-4.04	-5.27	1.98	1.14	-1.06	-0.64	-2.09	-1.18	2.13	0.75
December	3.34	4.58	6.35	5.14	-10.15	-9.95	-5.95	-6.54	-1.98	0.51	-1.93	-3.67



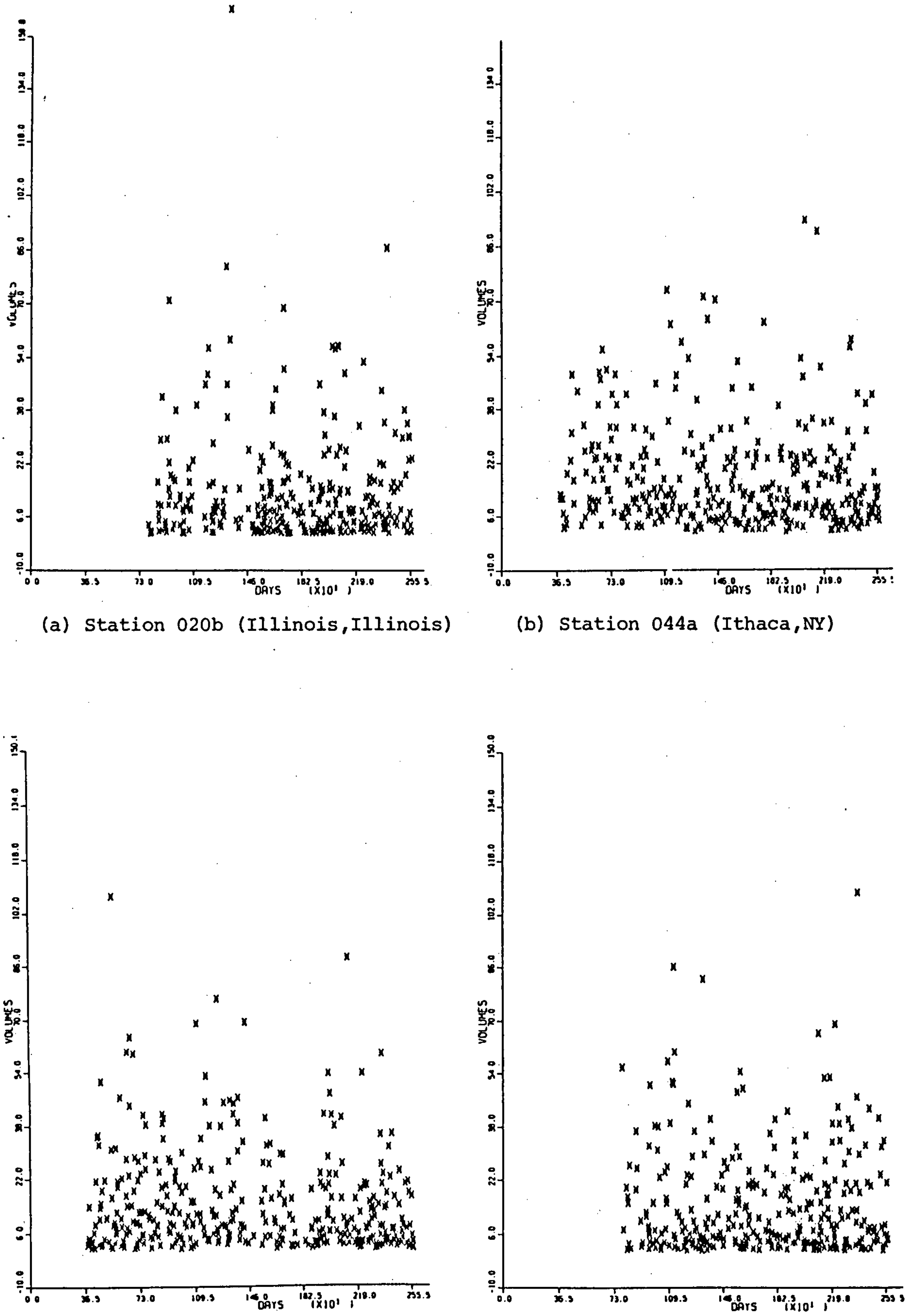
(a) Station 020b (Illinois,Illinois)

(b) Station 044a (Ithaca,NY)

(c) Station 048a (Brookhaven,NY)

(d) Station 065a (Penn State,PA)

Figure 1.1; Field pH.



(a) Station 020b (Illinois,Illinois)

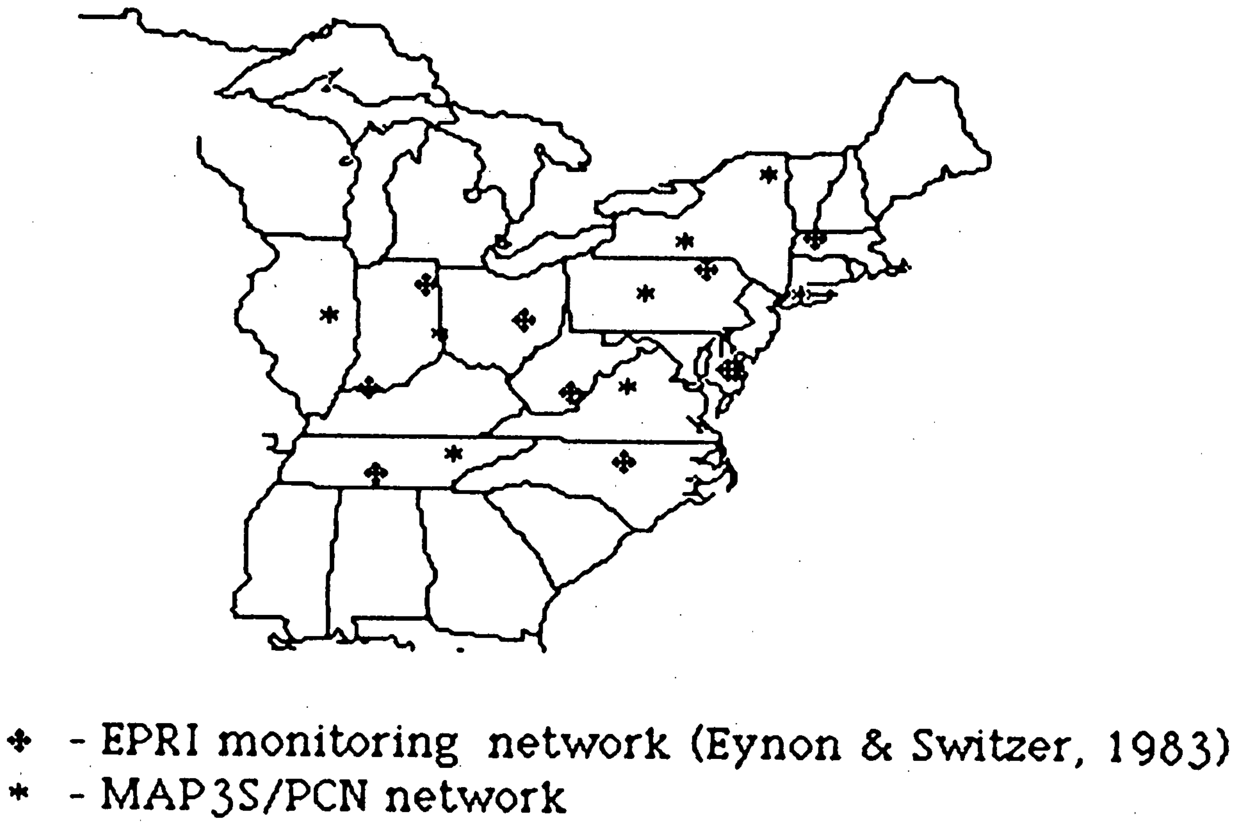
(b) Station 044a (Ithaca,NY)

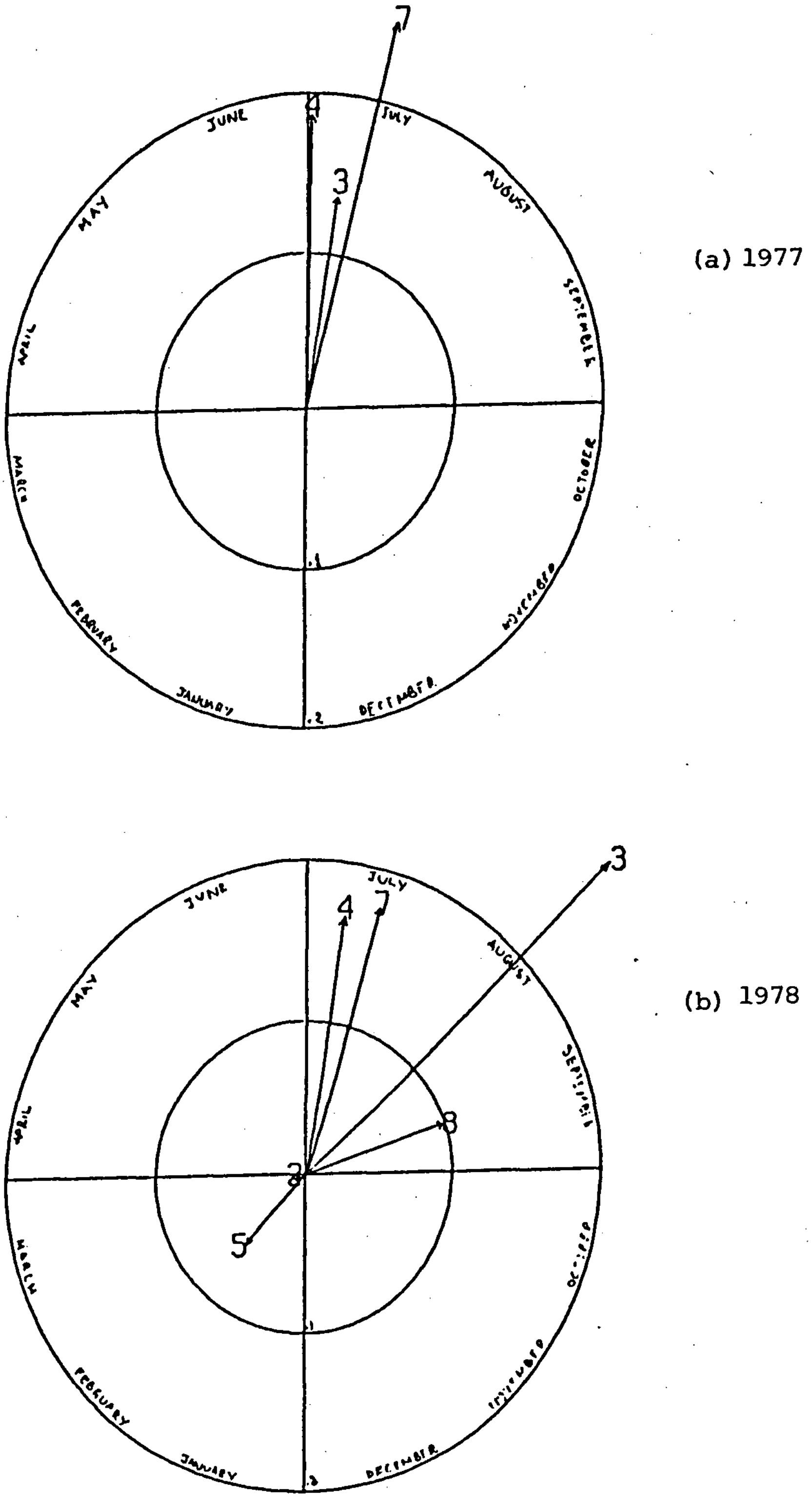
(c) Station 048a (Brookhaven,NY)

(d) Station 065a (Penn State,PA)

Figure 1.2: Rainfall volumes.

Figure 2.1 : The pH monitoring networks.



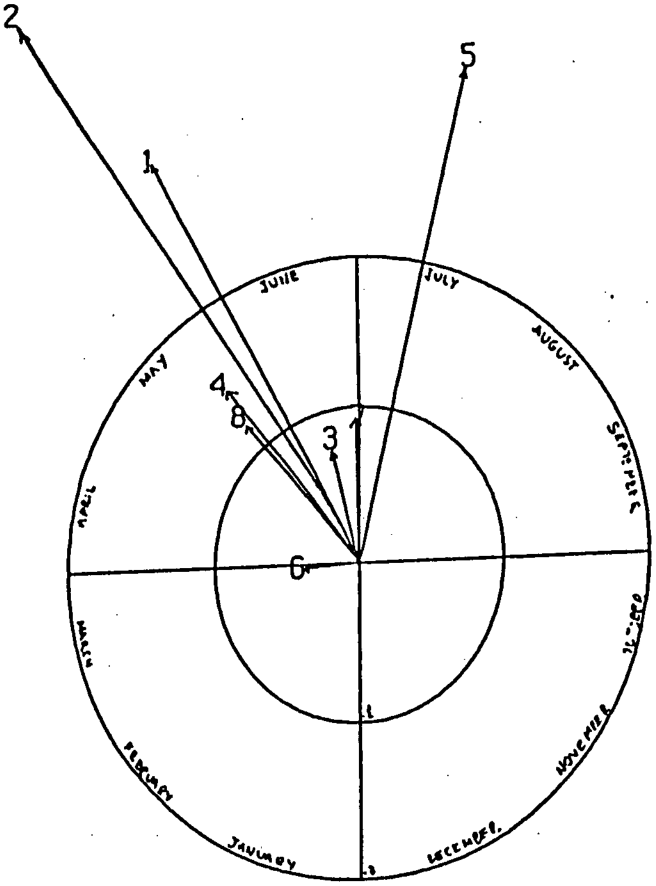


(a) 1977

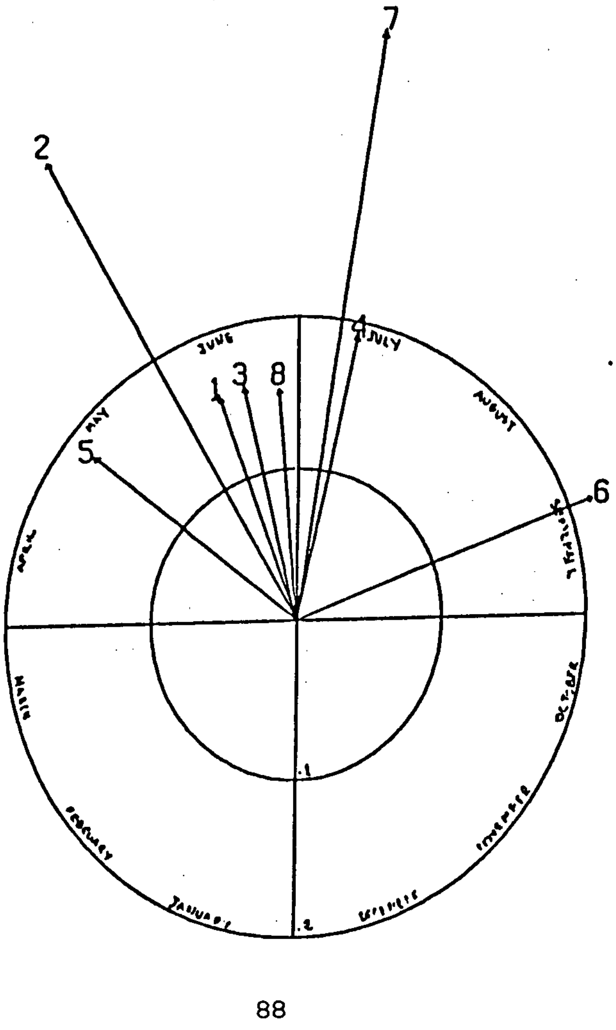
(b) 1978

Figure 2.2:

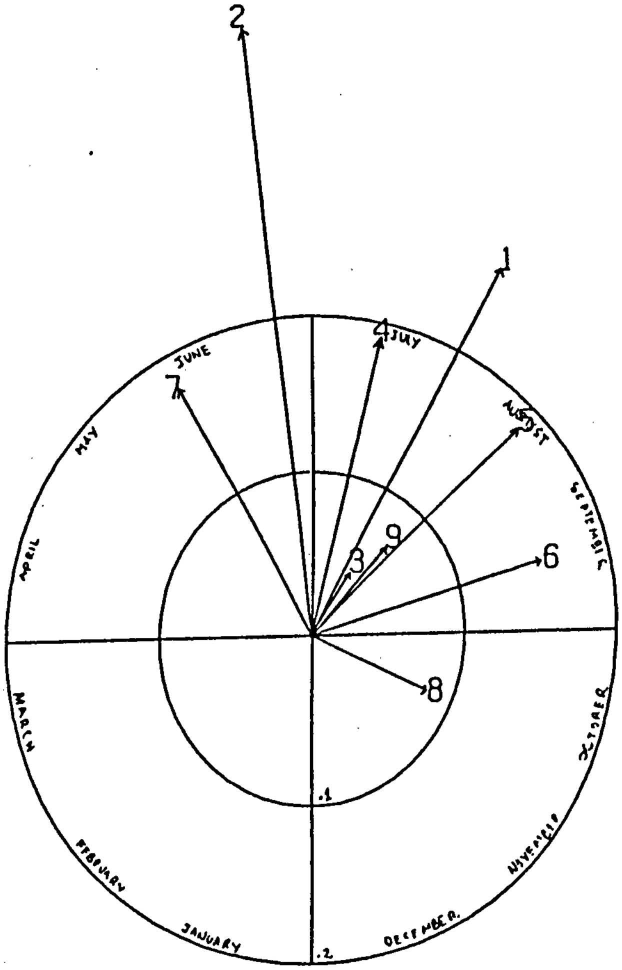
Phase and amplitude of seasonal pH variation: (a) 1977; (b) 1978; (c) 1979; (d) 1980; (e) 1981; (f) 1982. The direction of the ray for each station indicates the time of the year of lowest pH. The length of the ray indicates the amplitude of the seasonal variation.


2
5
1
4
8
3
7
6
JUNE
JULY
AUGUST
SEPTEMBER
OCTOBER
NOVEMBER
DECEMBER
JANUARY
FEBRUARY
MARCH
APRIL
MAY
.1
.2

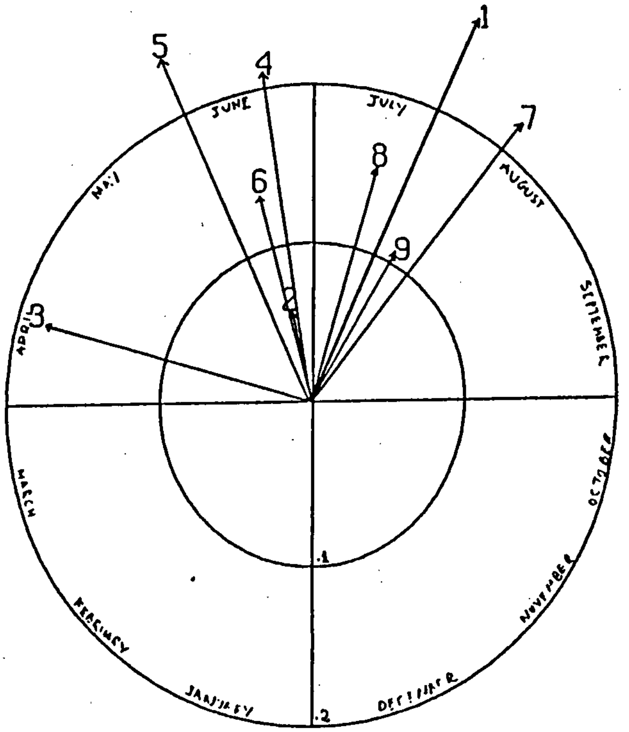
(c) 1979


7
2
4
1
3
8
5
6
JUNE
JULY
AUGUST
SEPTEMBER
OCTOBER
NOVEMBER
DECEMBER
JANUARY
FEBRUARY
MARCH
APRIL
MAY
.1
.2

(d) 1980


2
1
4
JULY
JUNE
7
AUGUST
5
MAY
SEPTEMBER
9
3
6
APRIL
8
MARCH
OCTOBER
.1
FEBRUARY
NOVEMBER
JANUARY
DECEMBER
.2

(e) 1981


1
5
4
7
JUNE
JULY
8
MAY
6
AUGUST
9
SEPTEMBER
2
APRIL
3
MARCH
OCTOBER
.1
FEBRUARY
NOVEMBER
JANUARY
DECEMBER
.2

(f) 1982

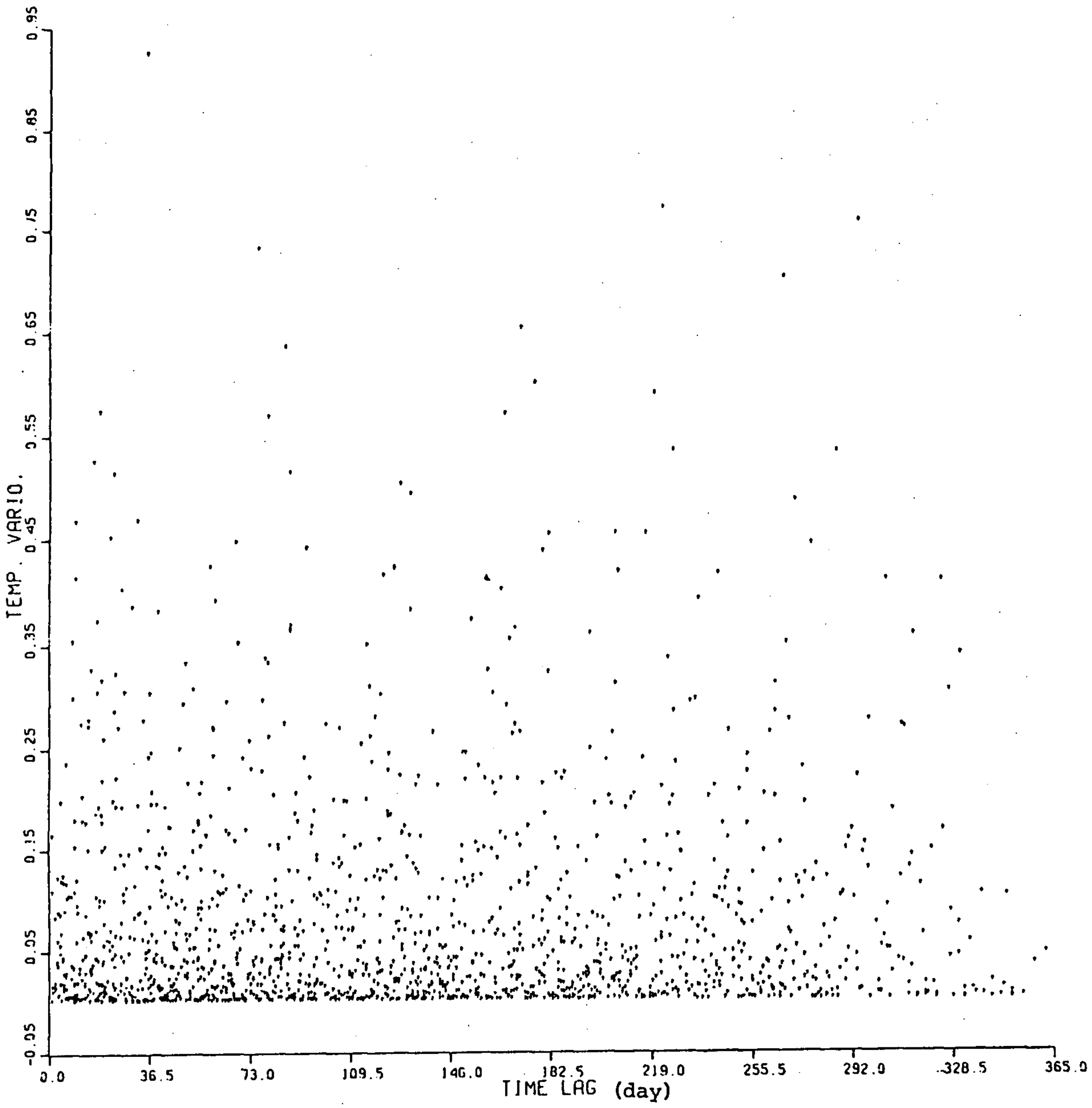


Figure 2.3: The estimates of the unsmoothed temporal variogram obtained for data of station 072a, in 1982.

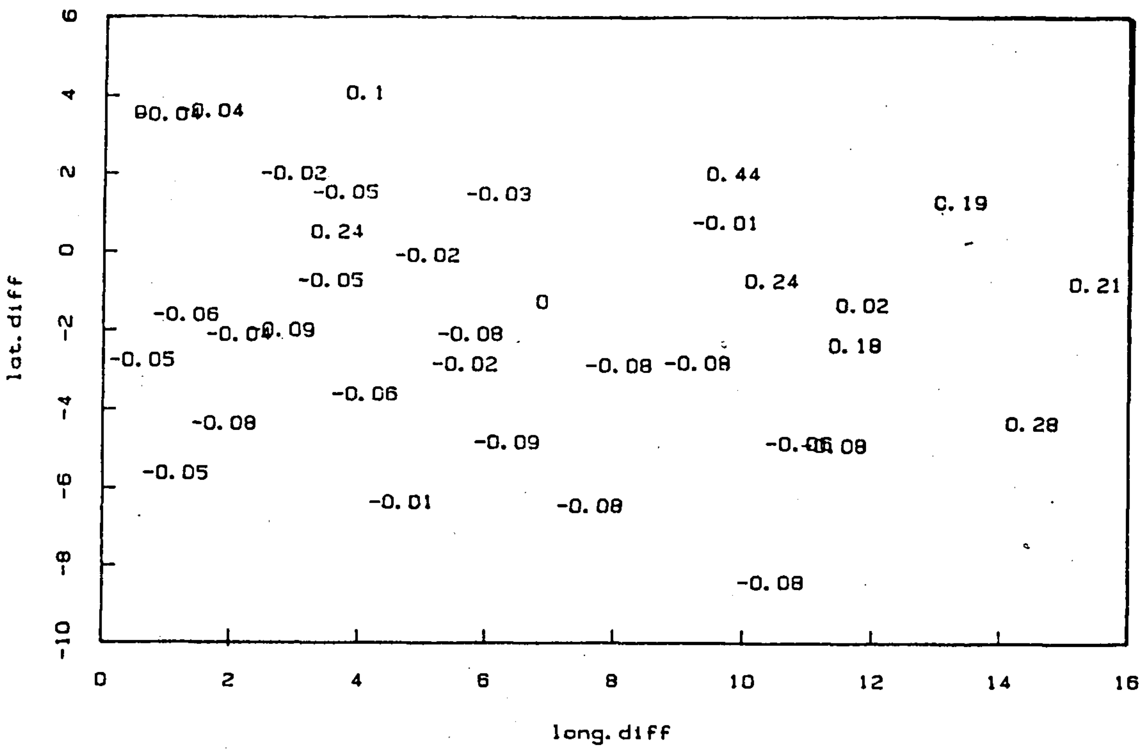


Figure 2.4 :
The estimates of the unsmoothed spatial variogram for different station pairs. LONGITUDE and LATITUDE (in degree) denote the difference in longitude and in latitude, respectively, between a station pair. All years of data combined.

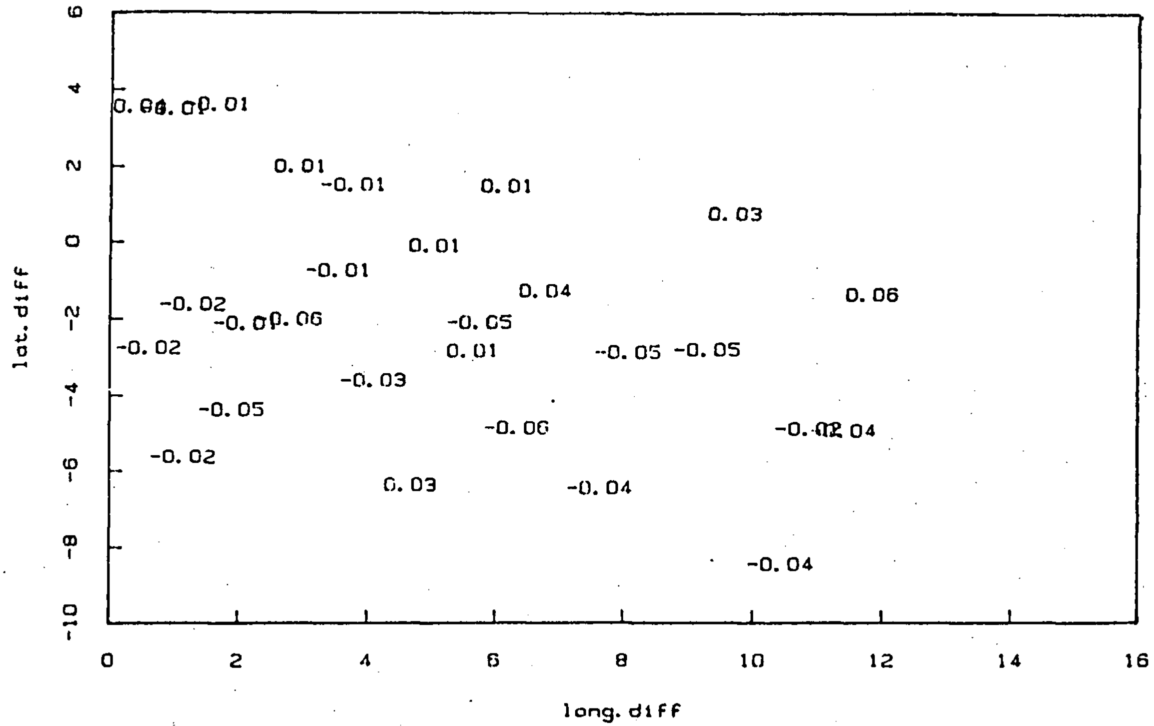
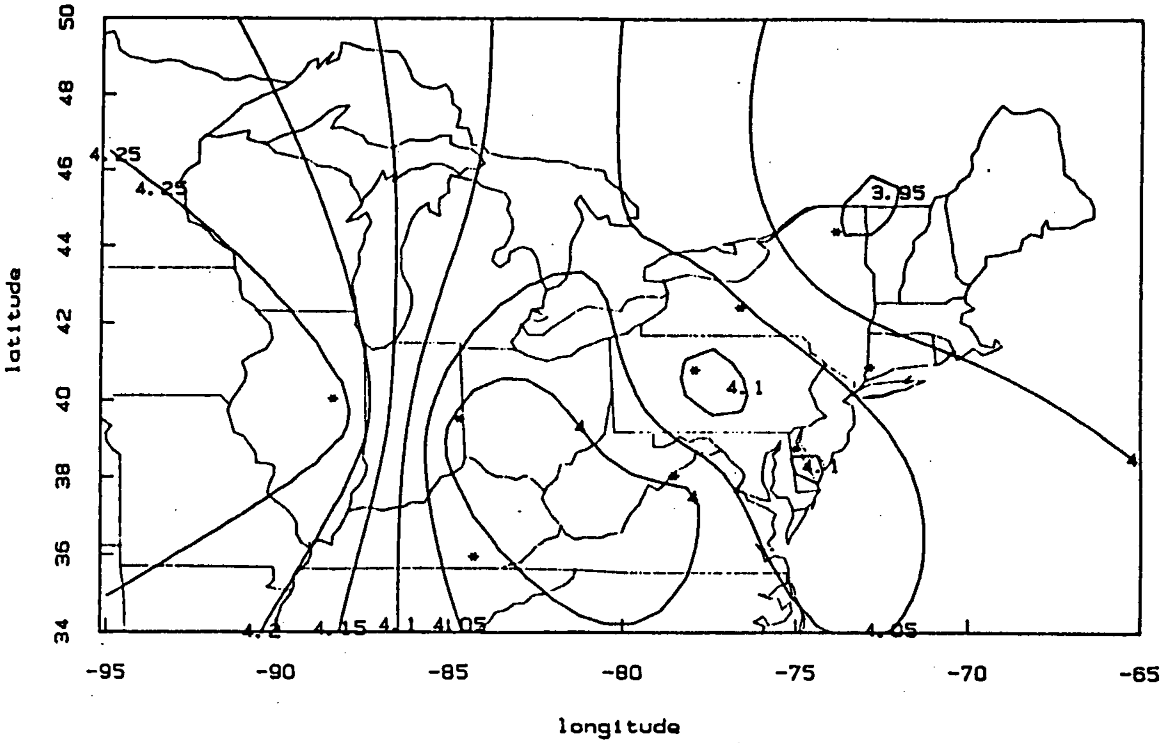
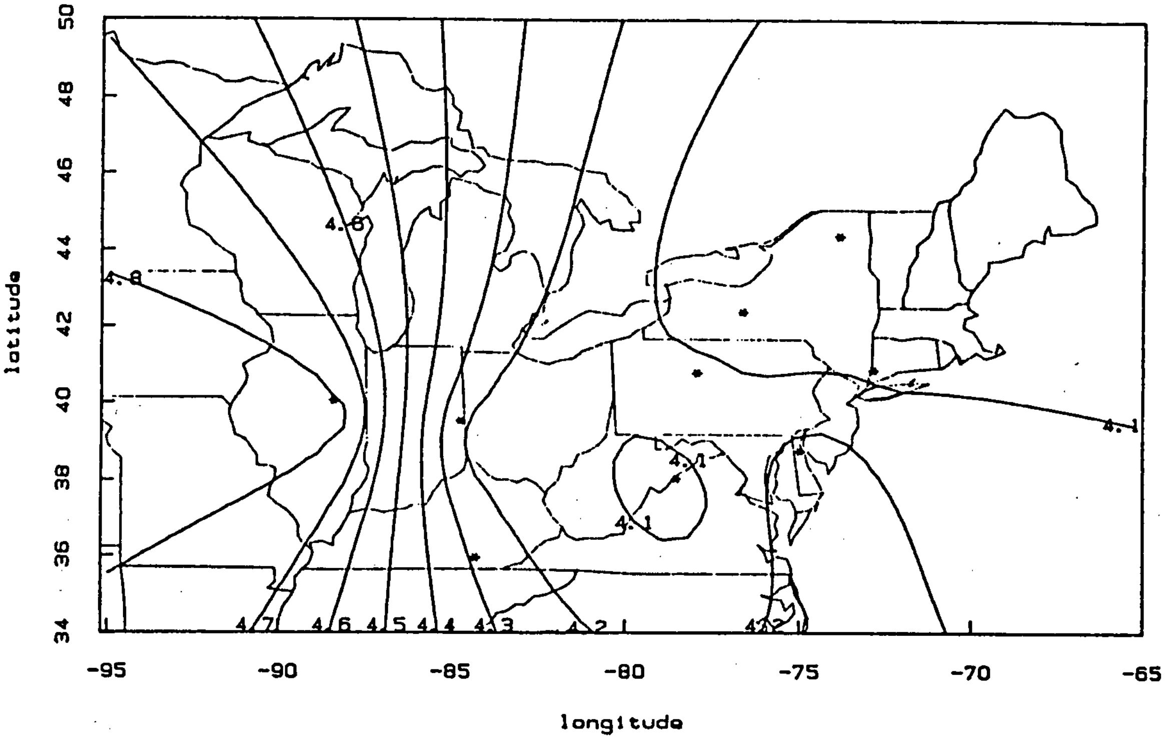


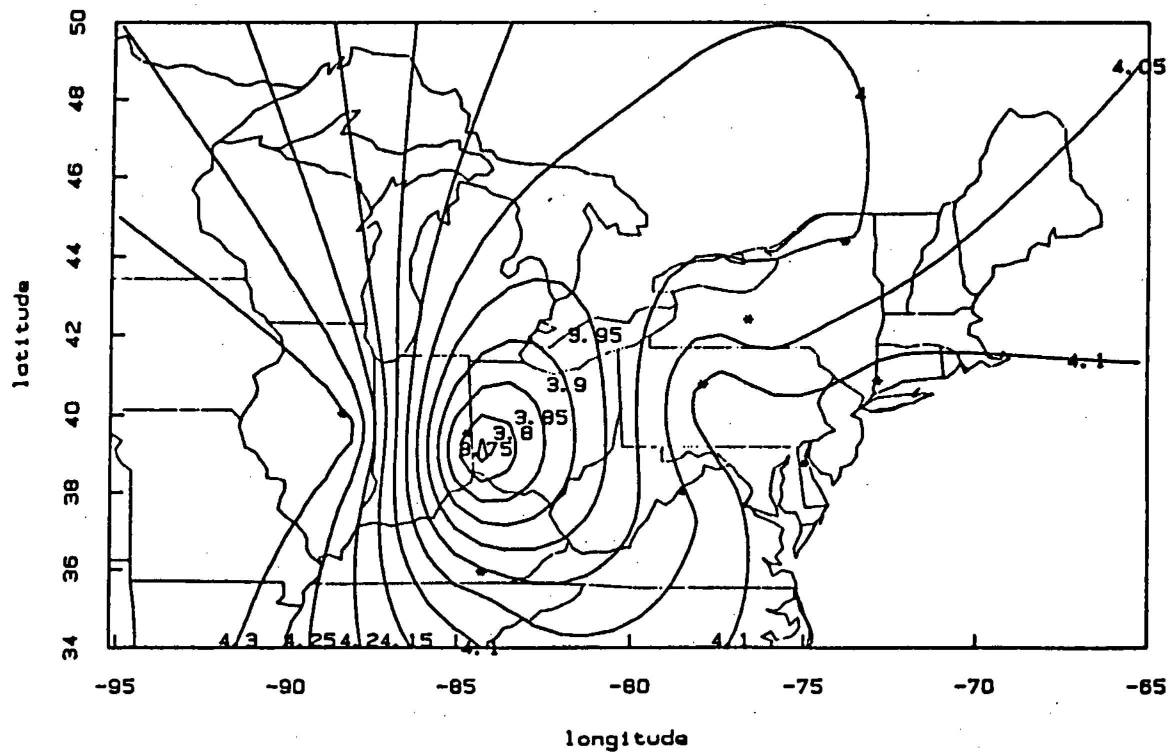
Figure 2.5 : Same as above, data from Station 020b removed.



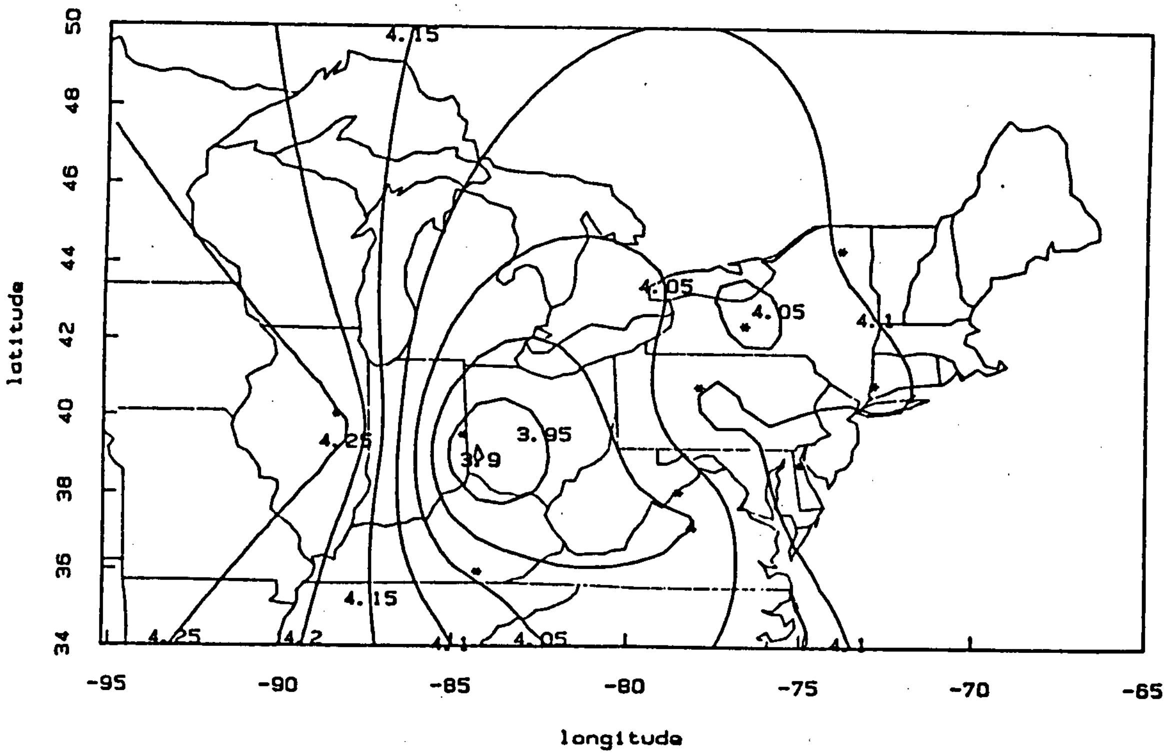
(a) Year 1979.



(b) Year 1980.



(c) Year 1981.



(d) Year 1982.

Figure 2.6 : Contour maps for pH.
'*': denotes the location of station.

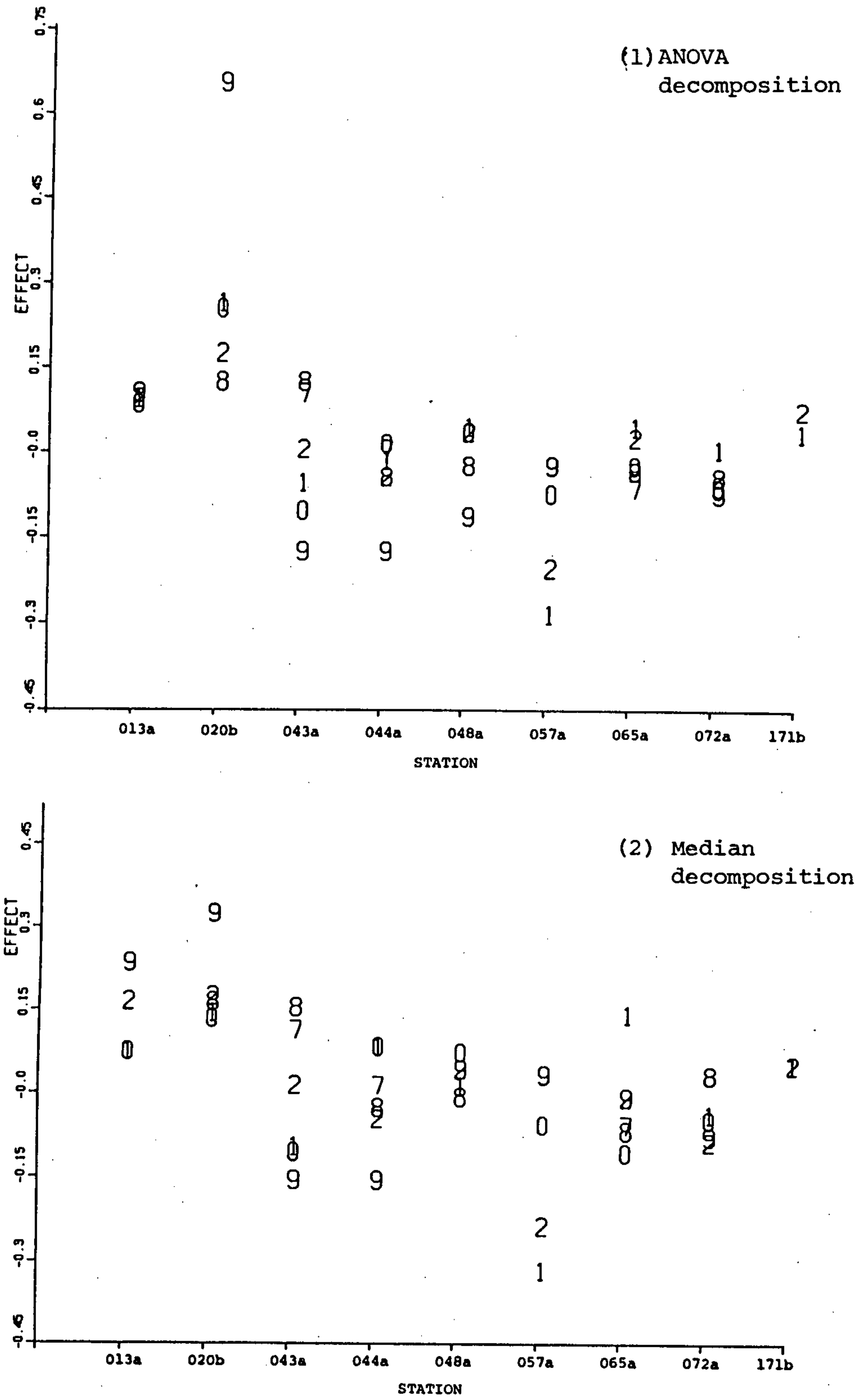
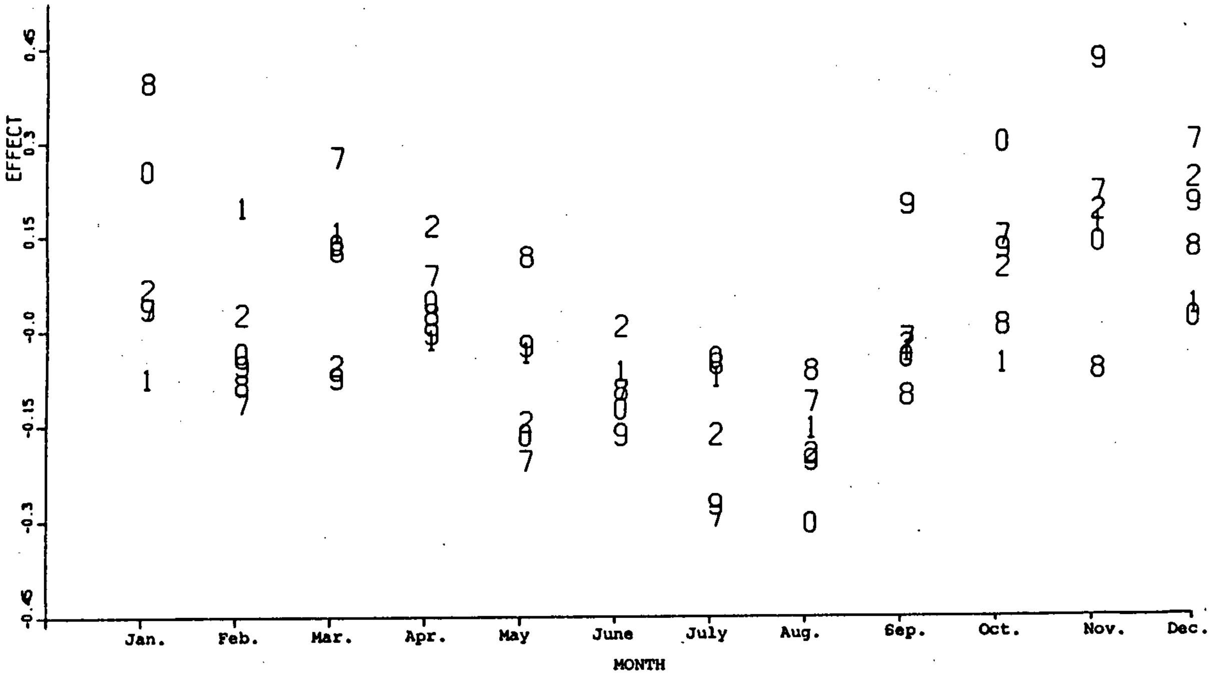


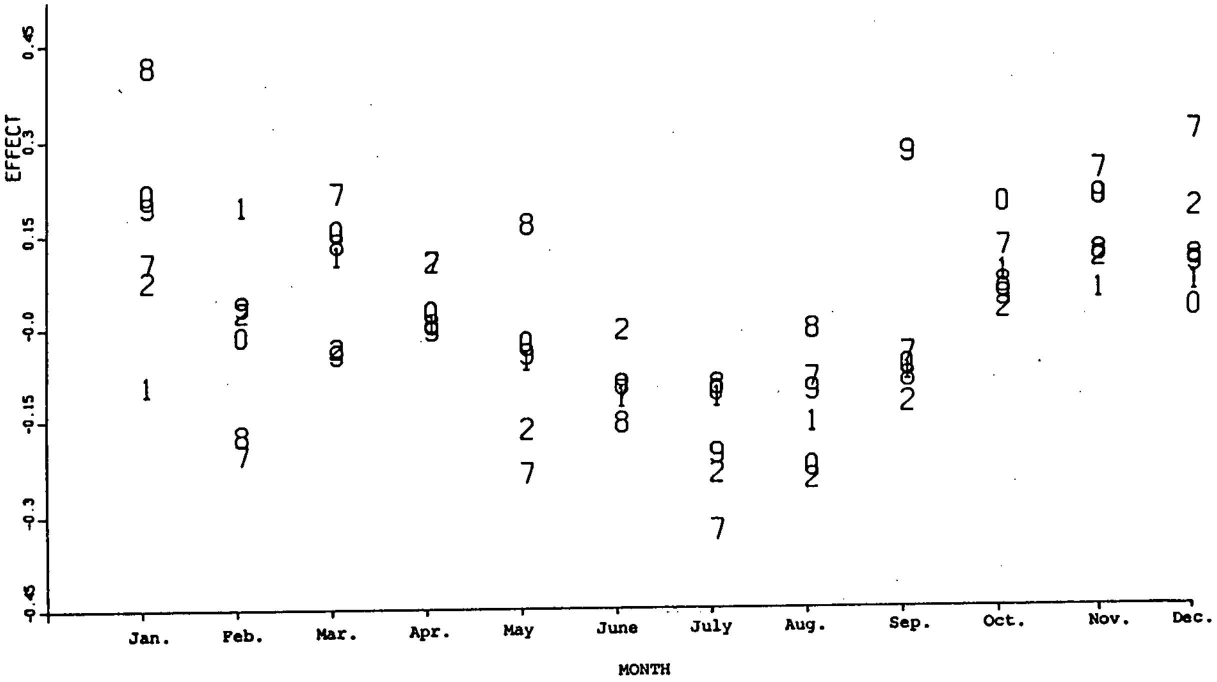
Figure 3.1: Station effects (pH levels).

Legend: 7=1977, 8=1978, 9=1979, 0=1980, 1=1981, 2=1982.



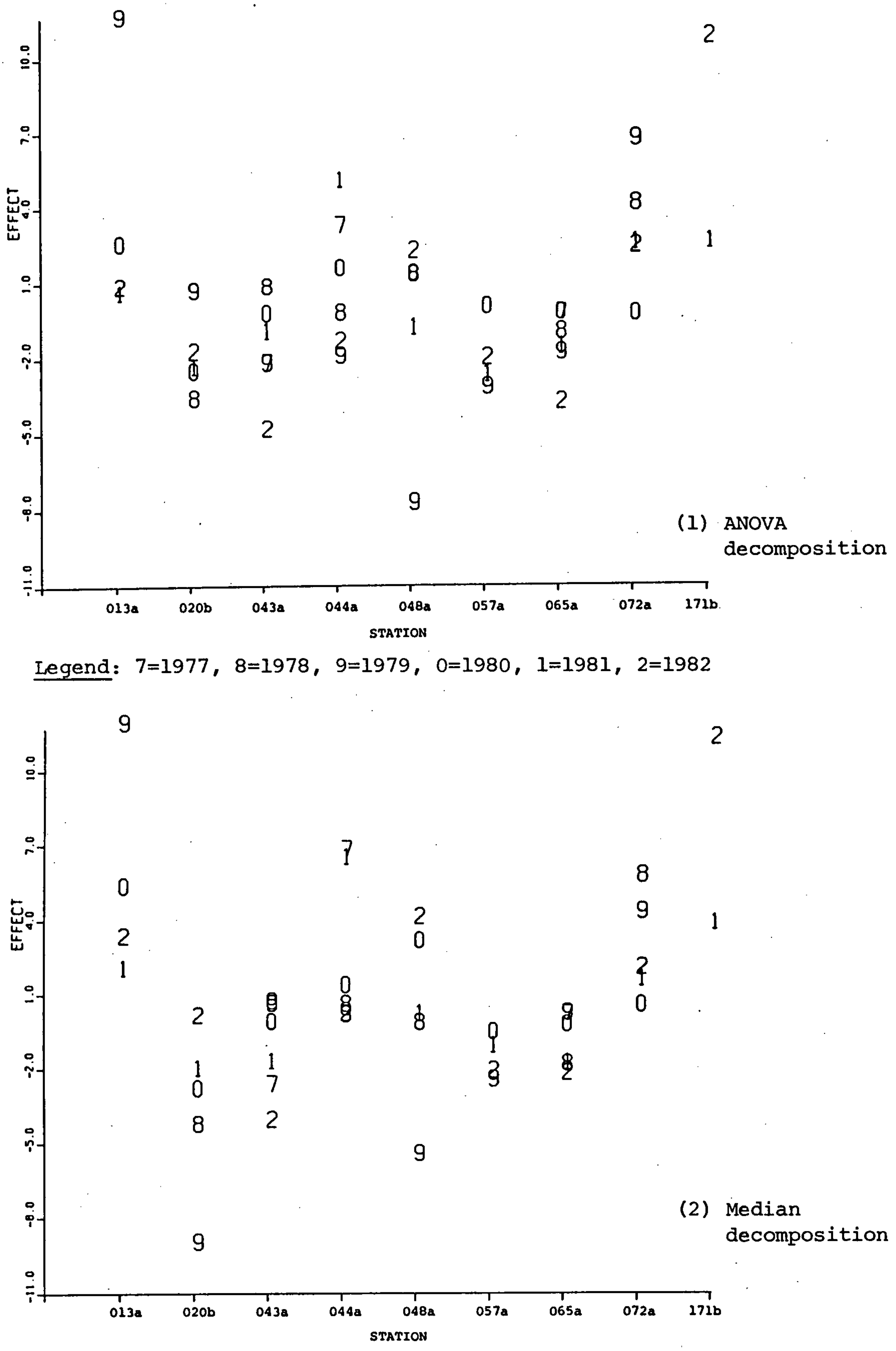
(1) ANOVA decomposition

Legend: 7=1977, 8=1978, 9=1979, 0=1980, 1=1981, 2=1982.



(2) Median decomposition

Figure 3.2: Monthly effects (pH levels).

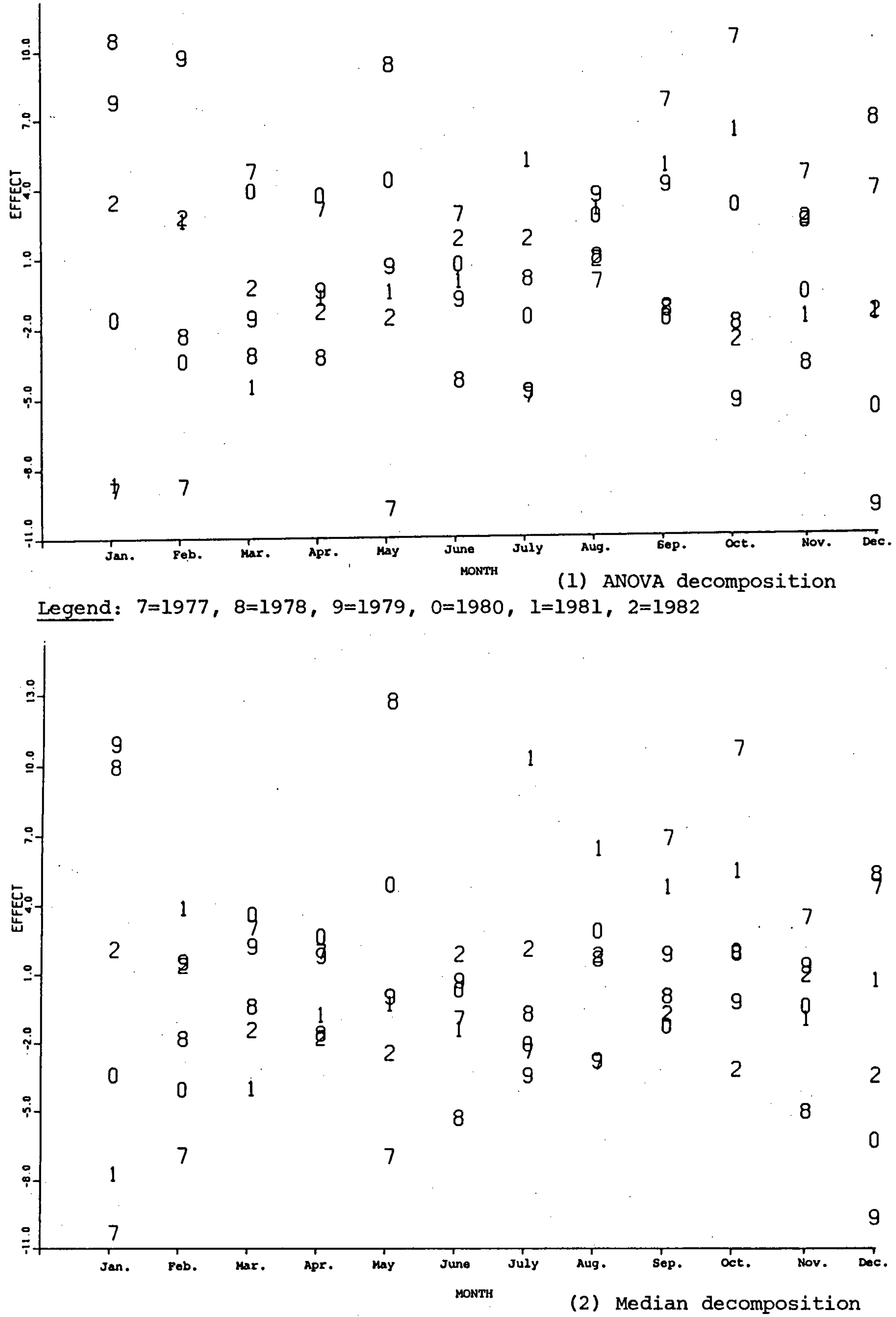


(1) ANOVA decomposition

Legend: 7=1977, 8=1978, 9=1979, 0=1980, 1=1981, 2=1982

(2) Median decomposition

Figure 3.3: Station effects (rainfall volumes)



(1) ANOVA decomposition

Legend: 7=1977, 8=1978, 9=1979, 0=1980, 1=1981, 2=1982

(2) Median decomposition

Figure 3.4: Monthly effects (rainfall volumes).

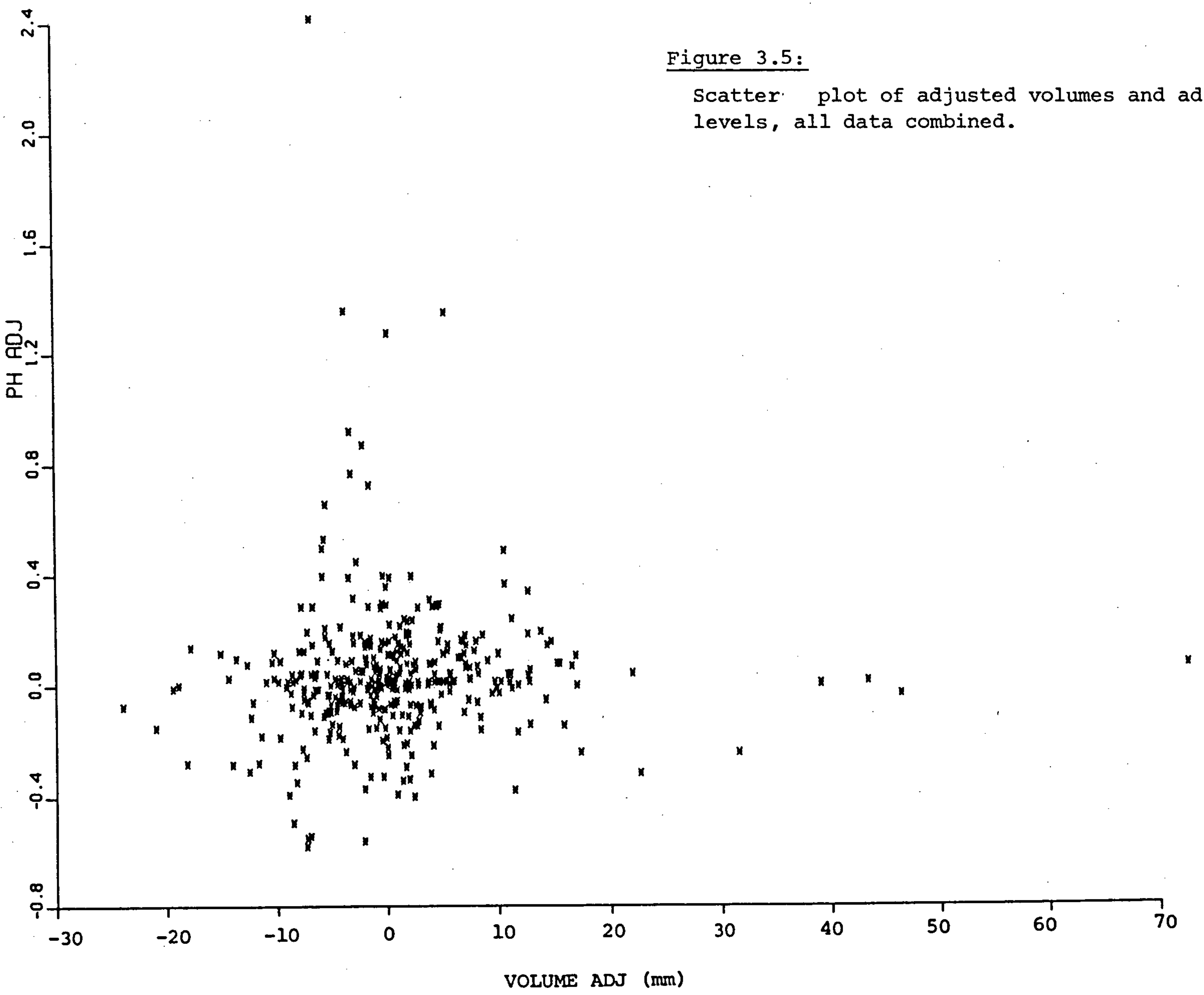
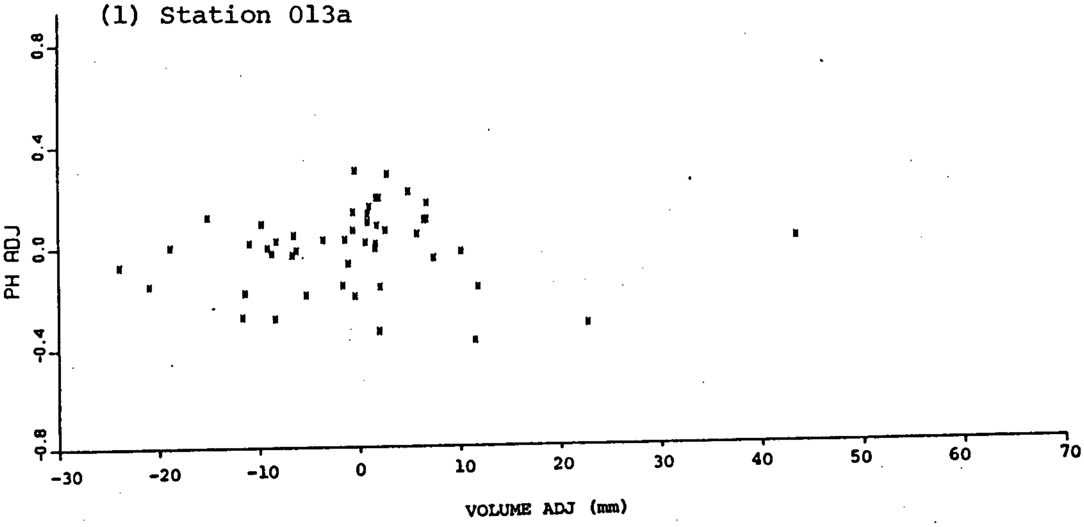


Figure 3.5:

Scatter plot of adjusted volumes and adjusted pH levels, all data combined.



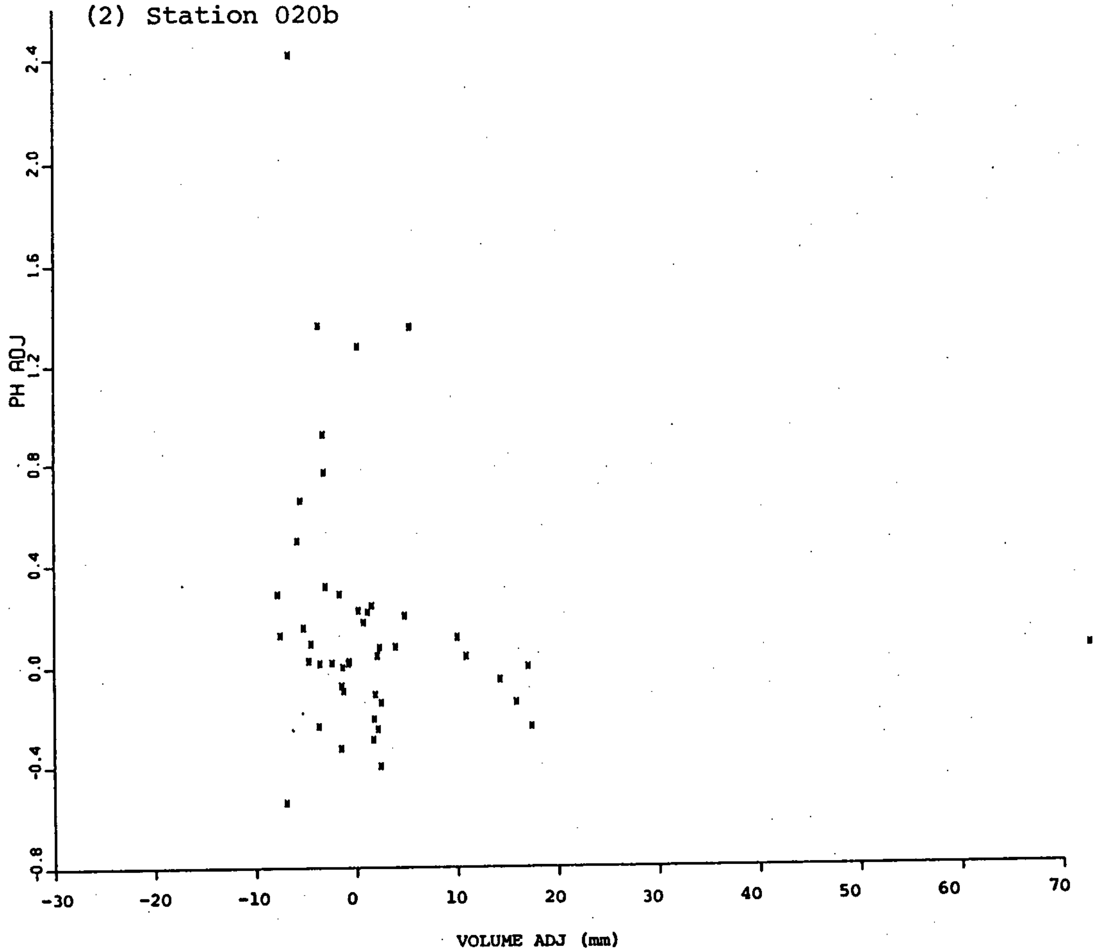
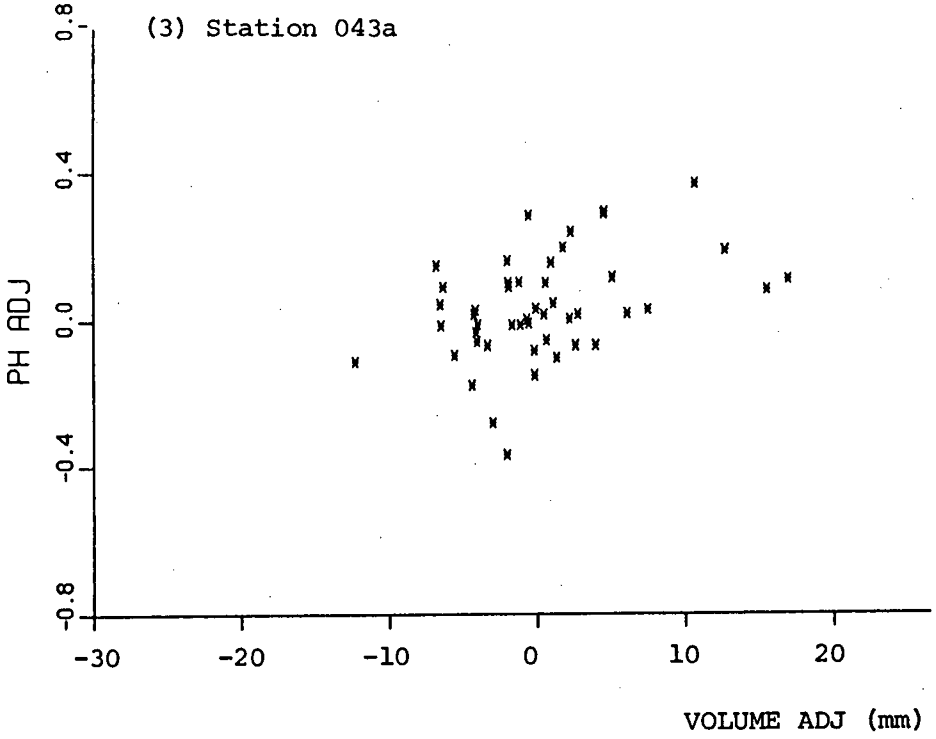
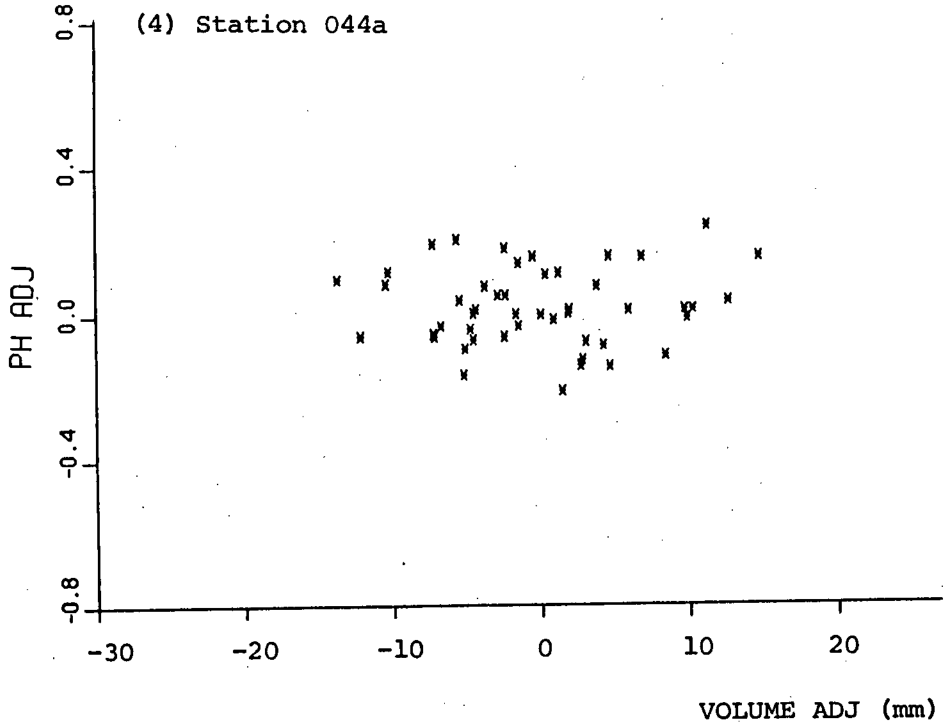
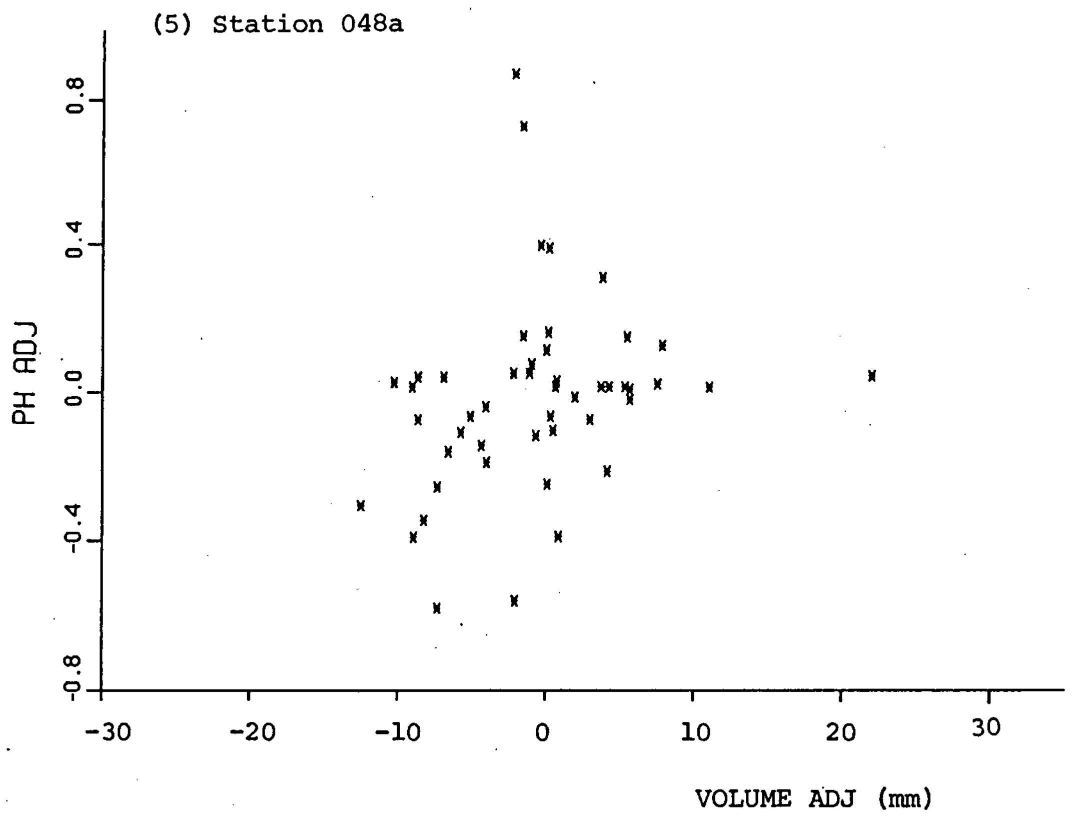
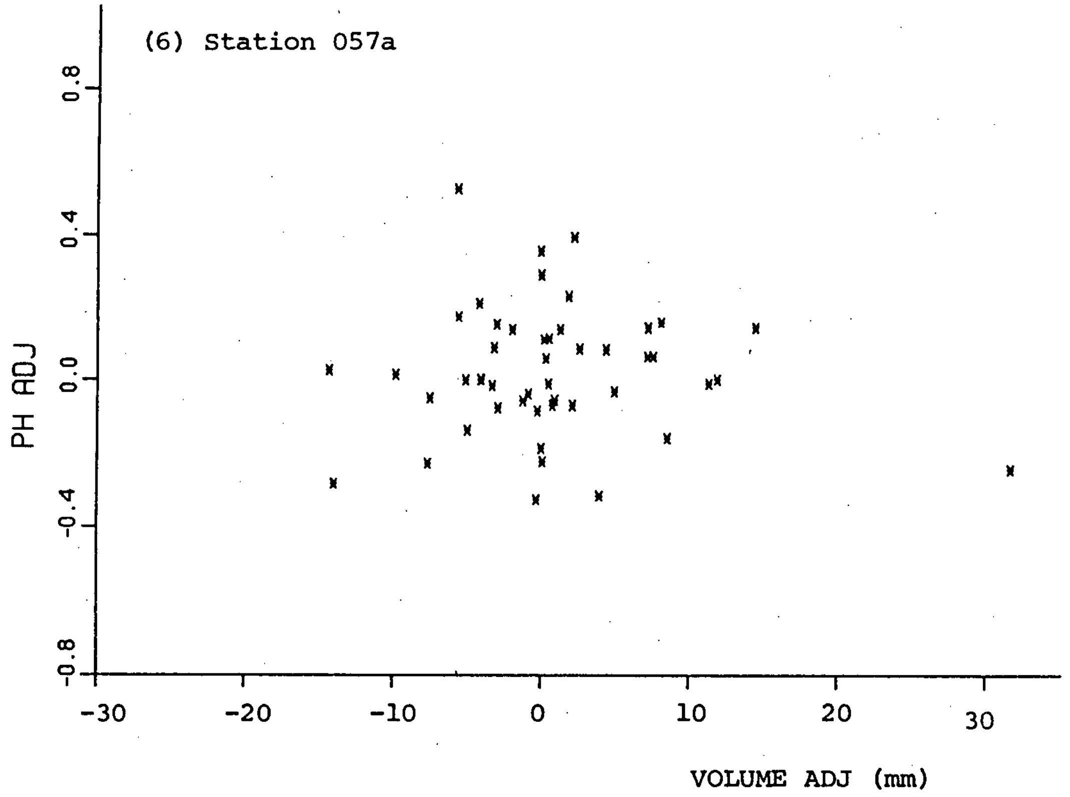


Figure 3.6: Scatter plots of adjusted volumes and adjusted pH levels for each station separately.

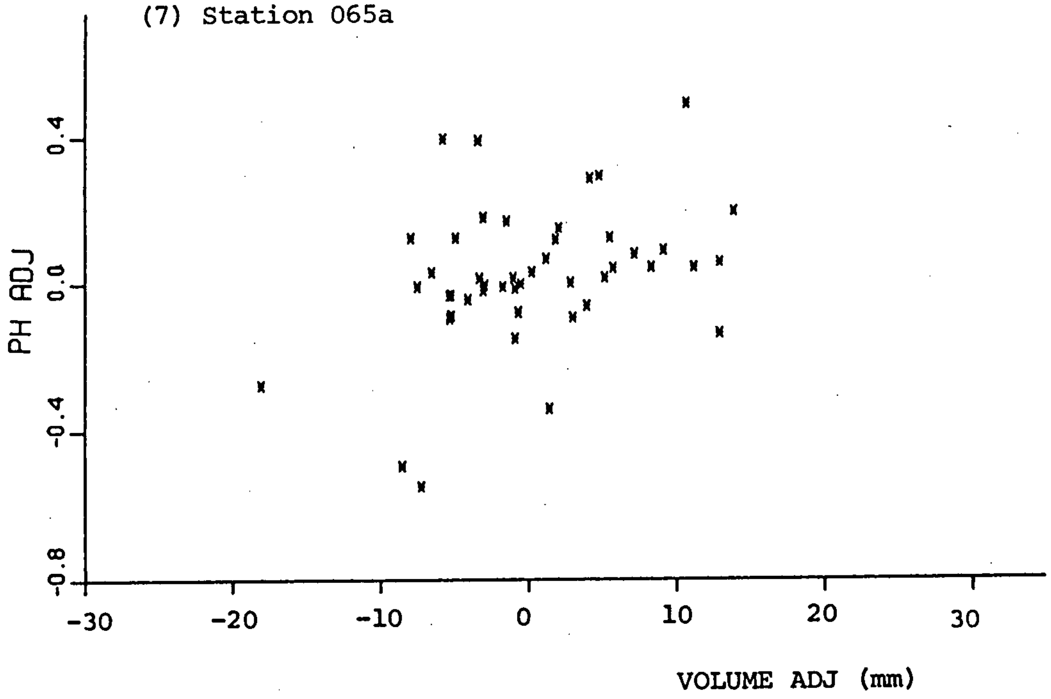

(3) Station 043a
PH ADJ
0.8
0.4
0.0
-0.4
-0.8
-30
-20
-10
0
10
20
VOLUME ADJ (mm)


(4) Station 044a
PH ADJ
0.8
0.4
0.0
-0.4
-0.8
-30
-20
-10
0
10
20
VOLUME ADJ (mm)

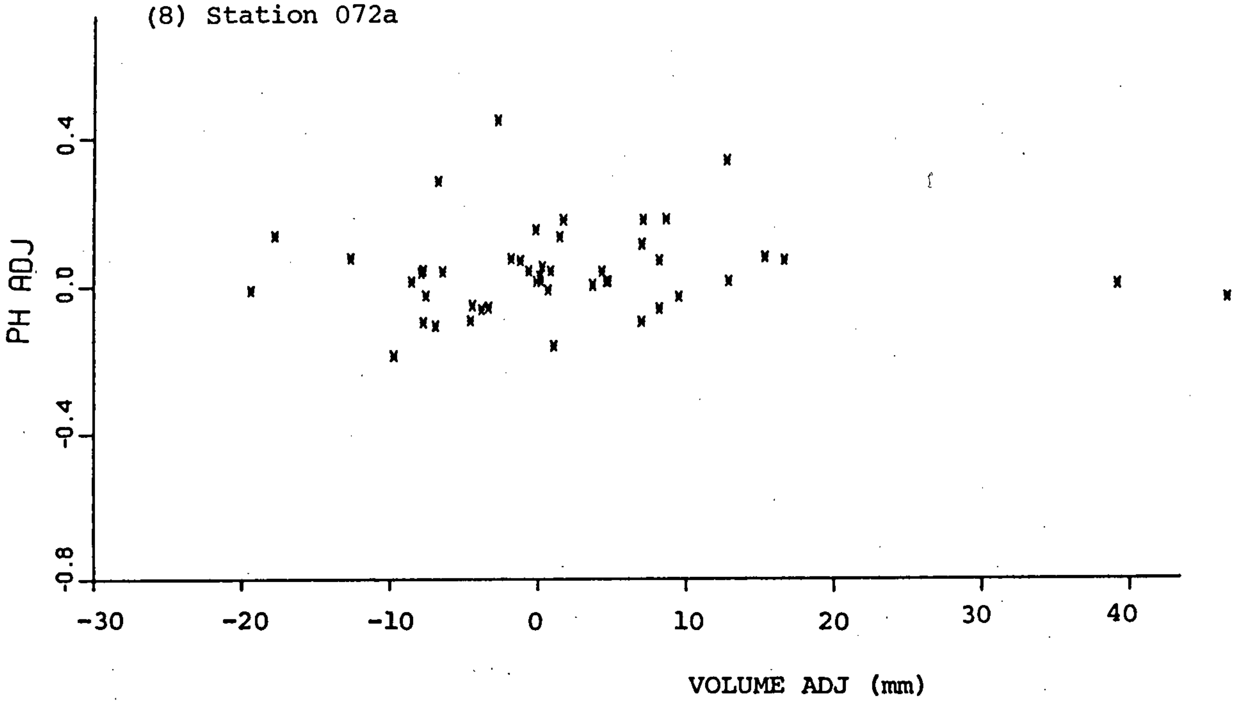

(5) Station 048a
PH ADJ
0.8
0.4
0.0
-0.4
-0.8
-30
-20
-10
0
10
20
30
VOLUME ADJ (mm)

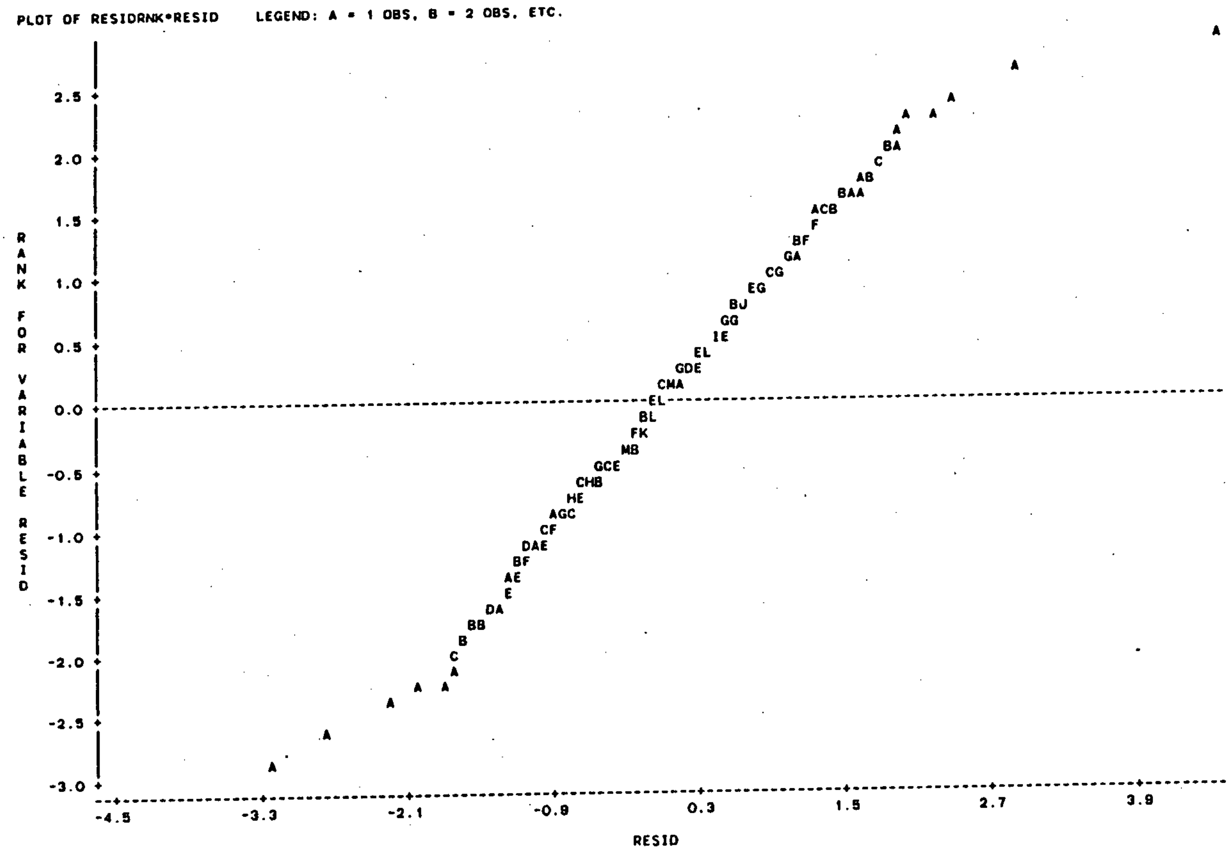

(6) Station 057a
PH ADJ
0.8
0.4
0.0
-0.4
-0.8
-30
-20
-10
0
10
20
30
VOLUME ADJ (mm)

(7) Station 065a

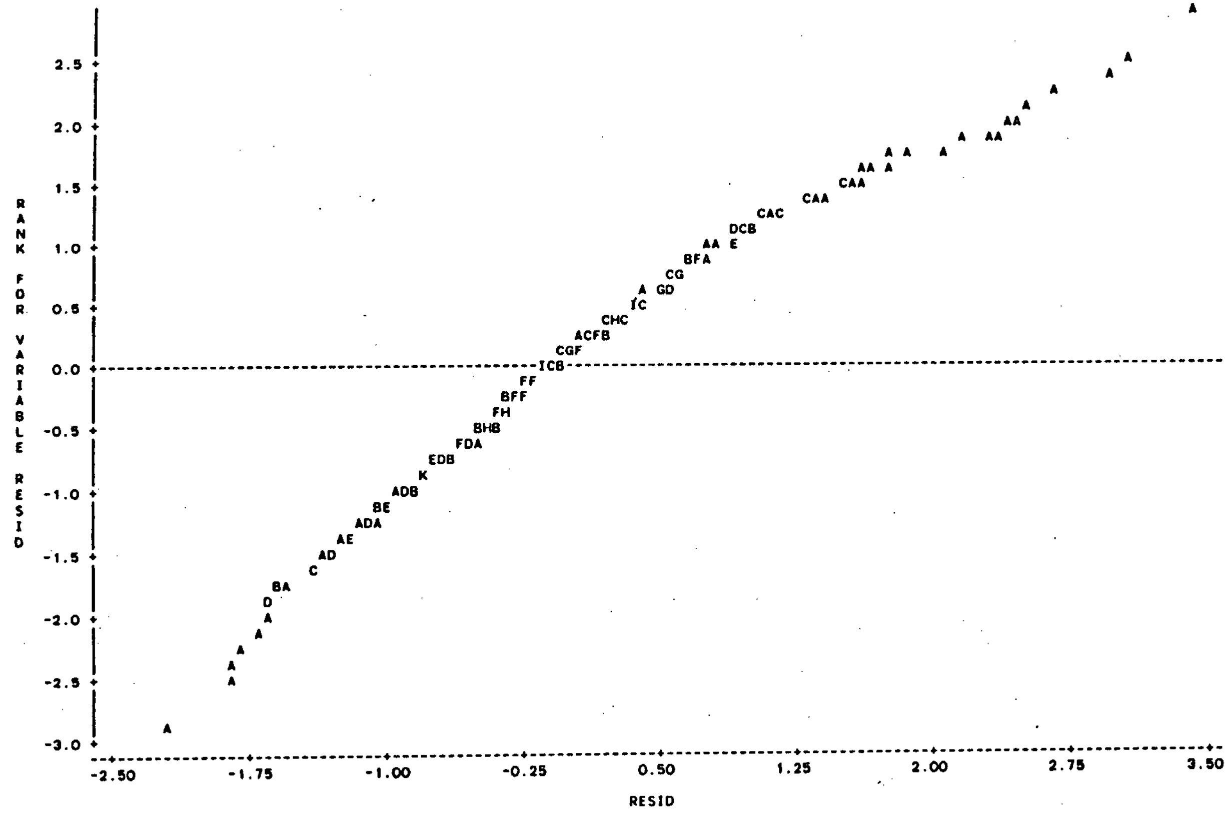


(8) Station 072a

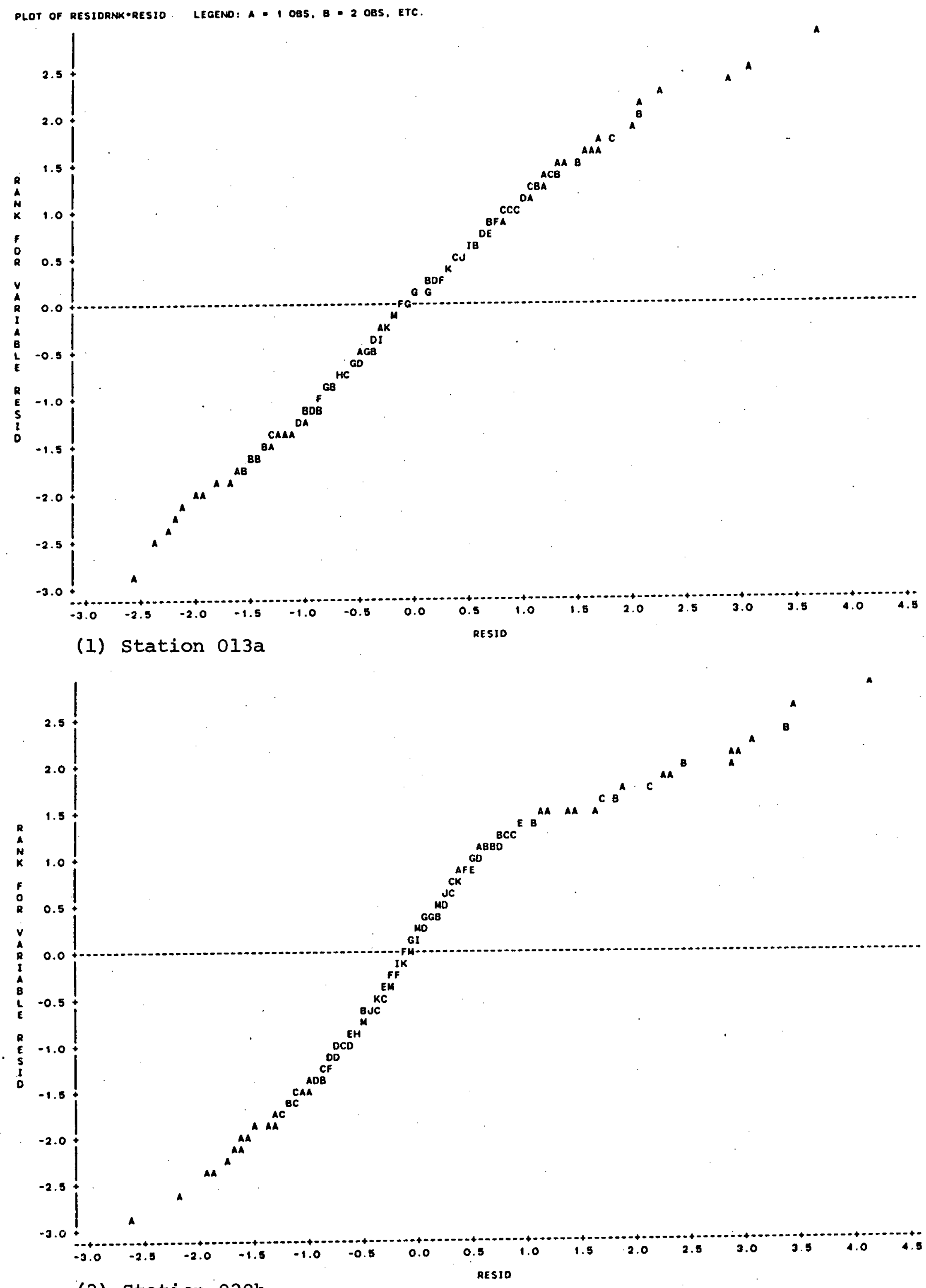




(5) Station 048a



(6) Station 057a

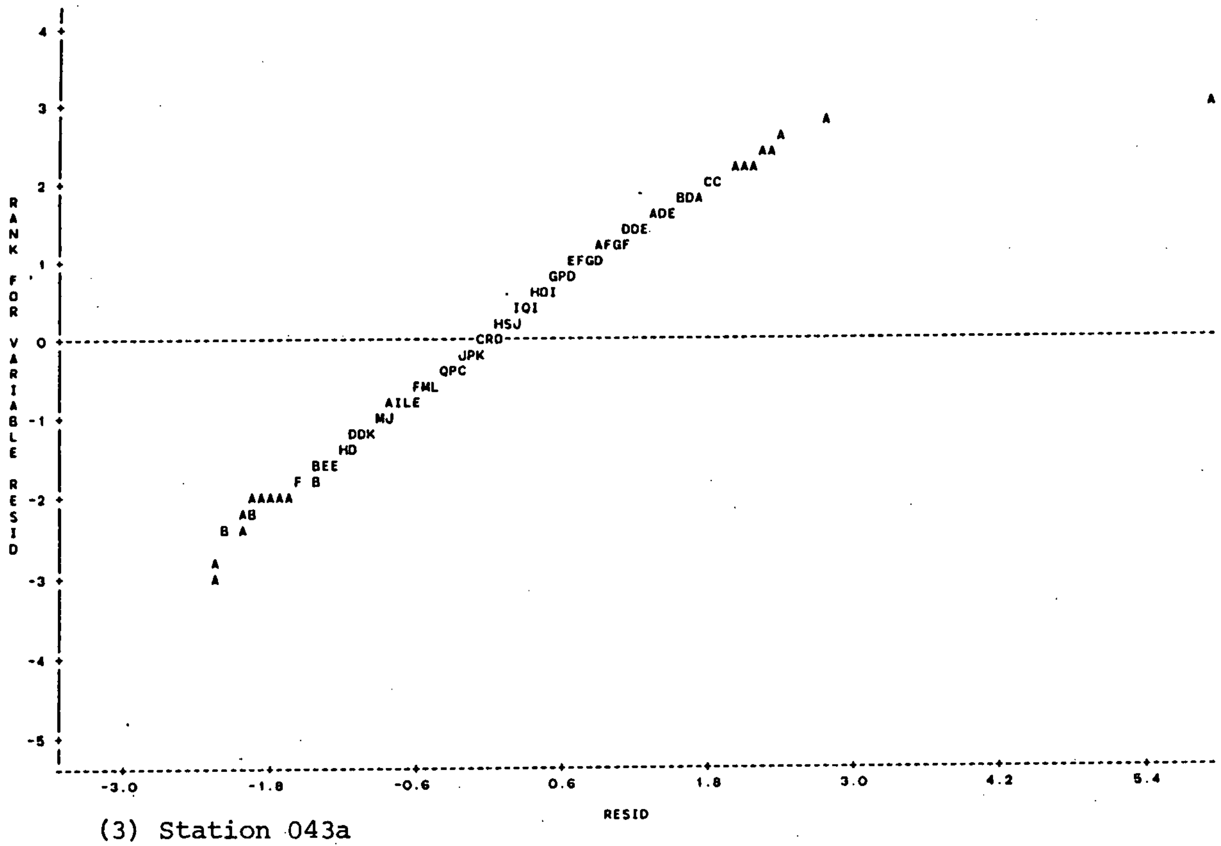


(1) Station 013a

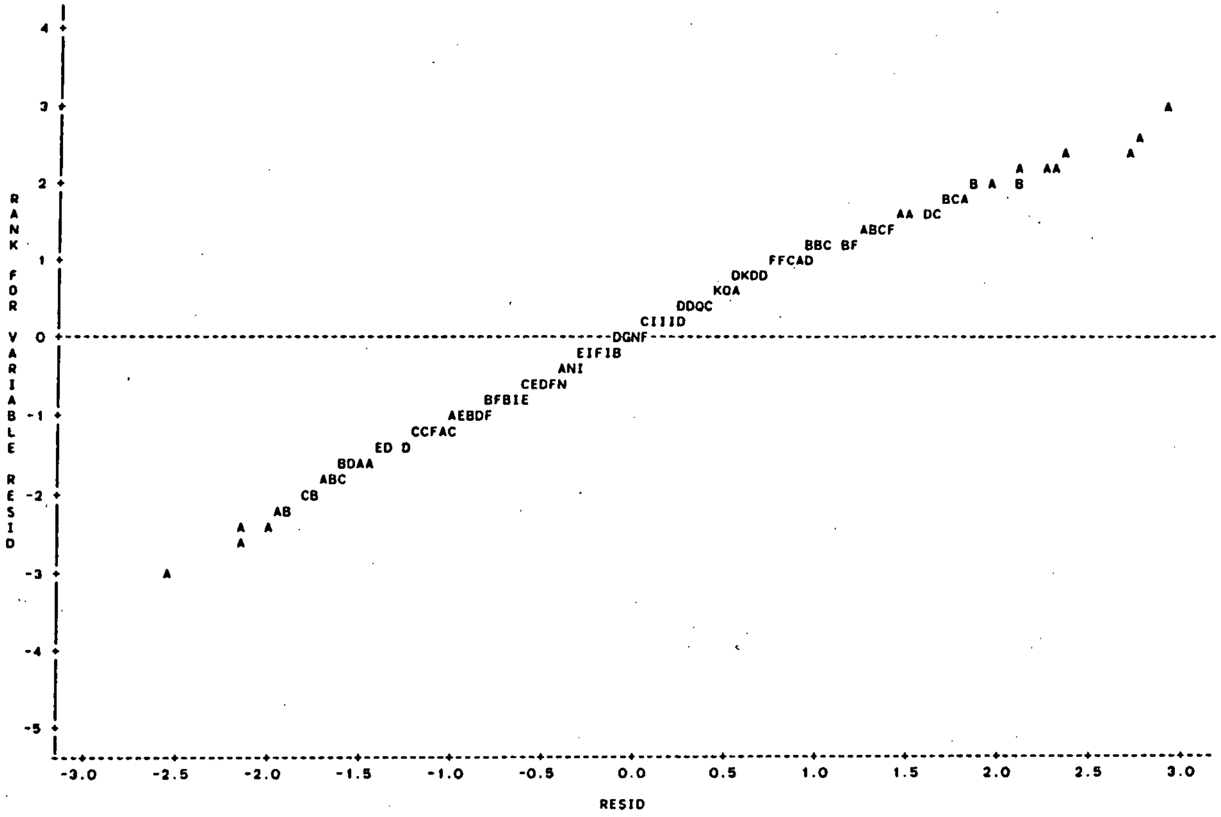
(2) Station 020b

Figure 3.7: Normal plots for each station (all years) separately.

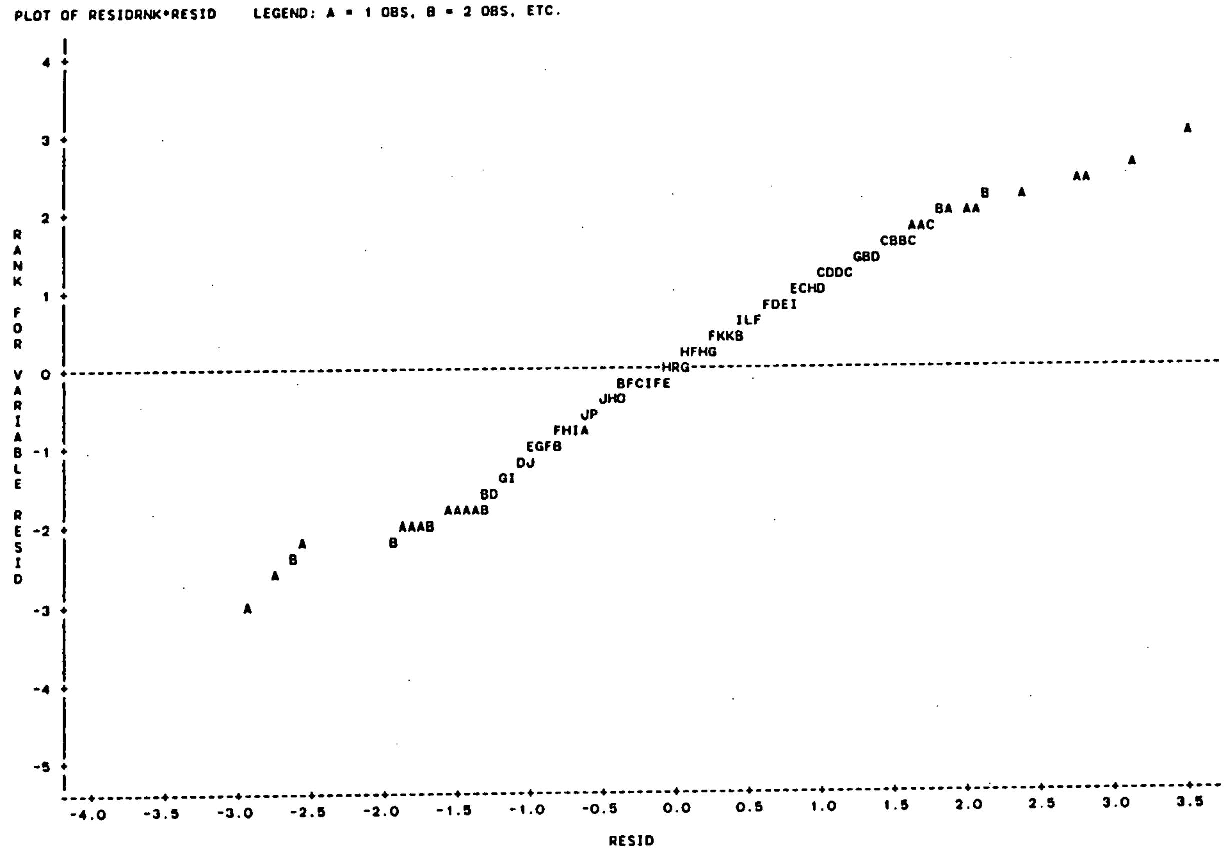
PLOT OF RESIDRNK*RESID   LEGEND: A = 1 OBS, B = 2 OBS, ETC.



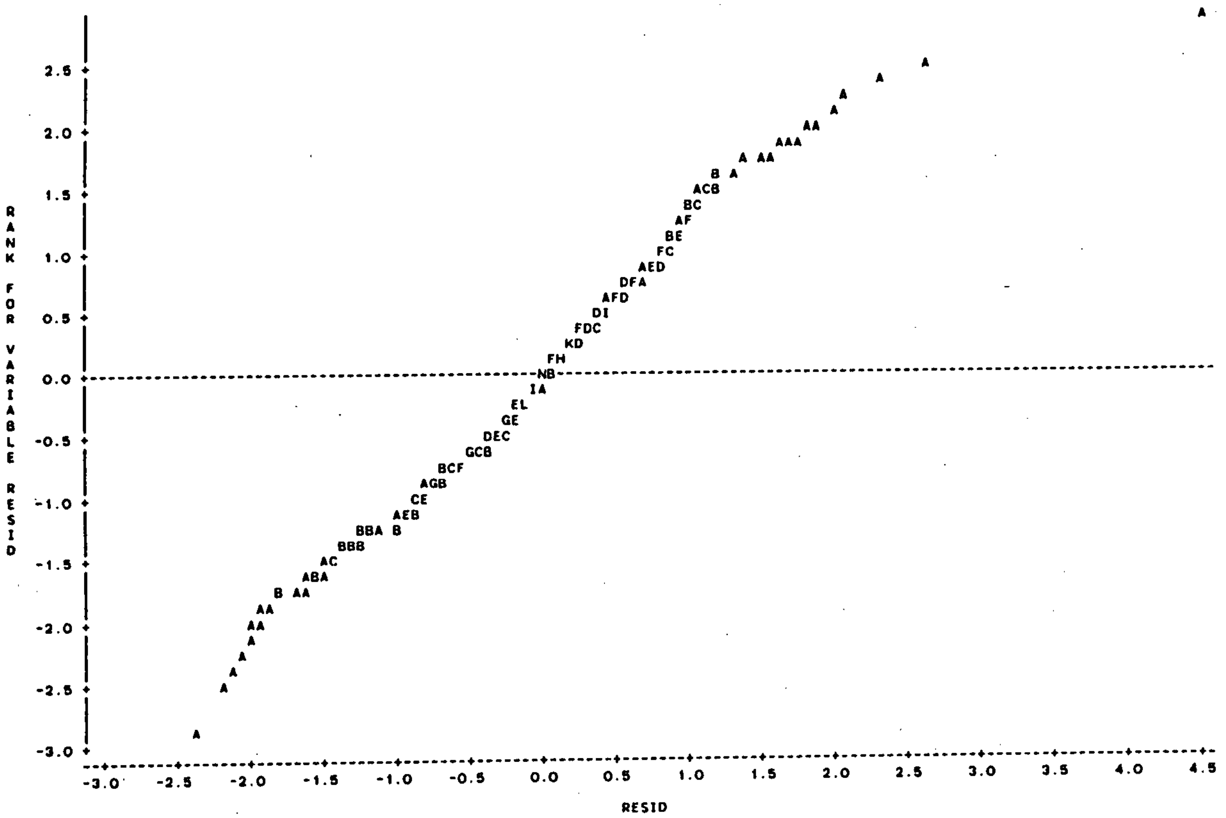
(3) Station 043a



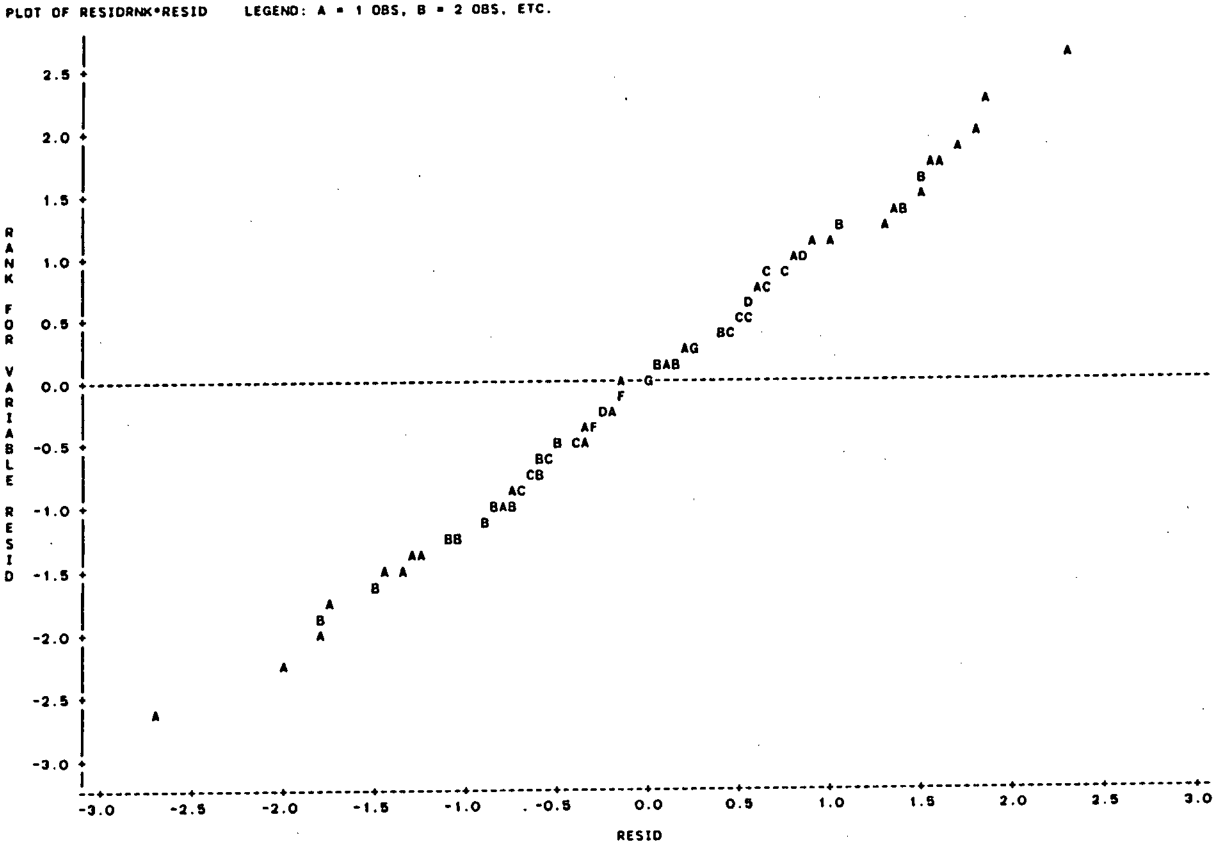
(4) Station 044a



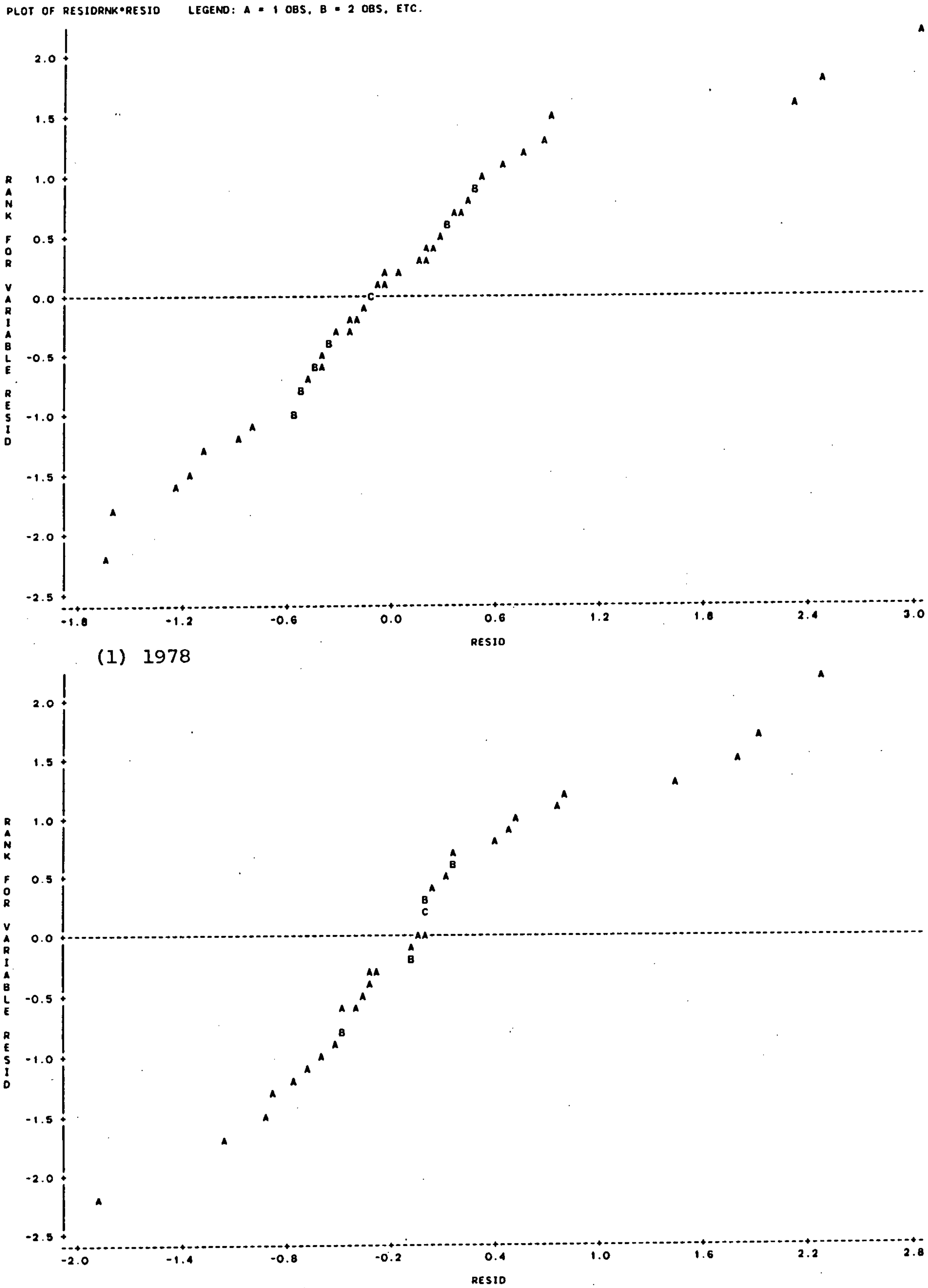
(7) Station 065a



(8) Station 072a



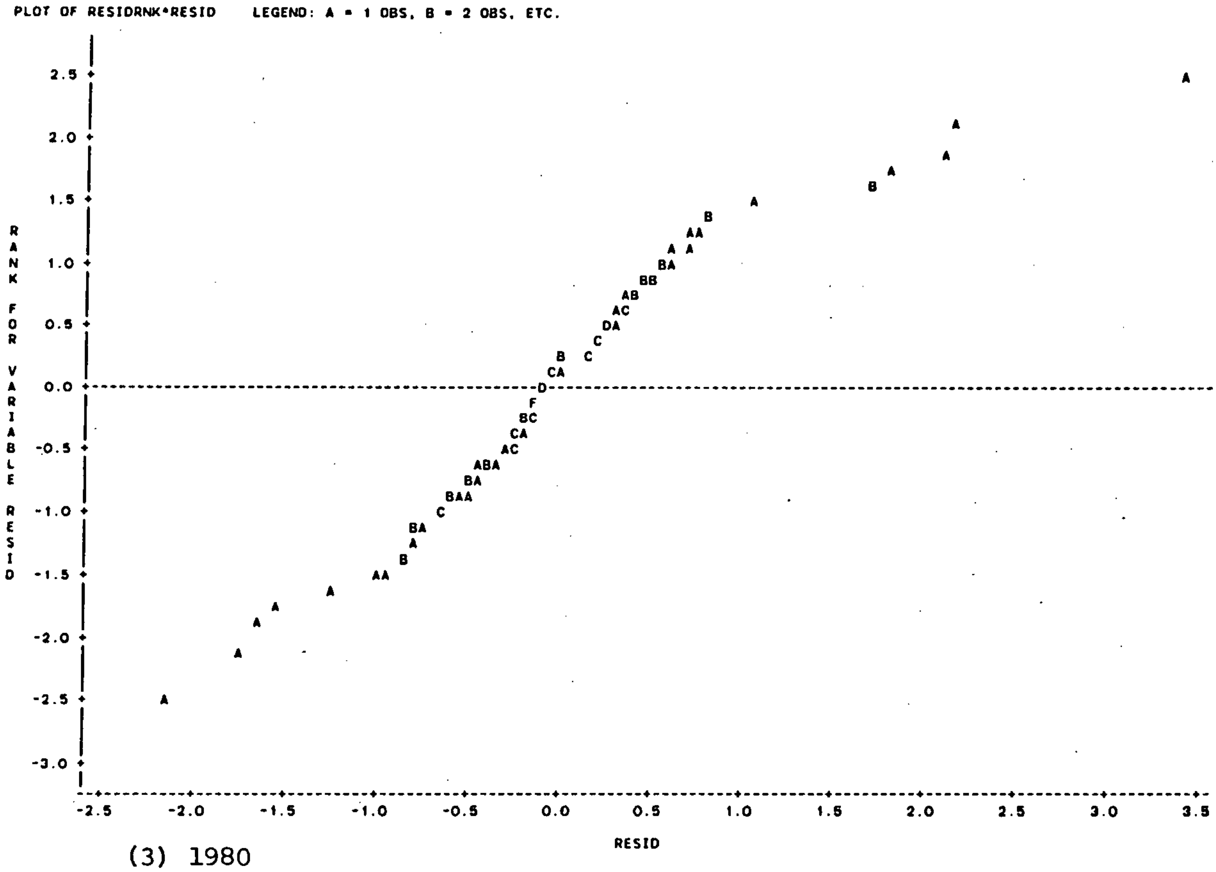
(9) Station 171b



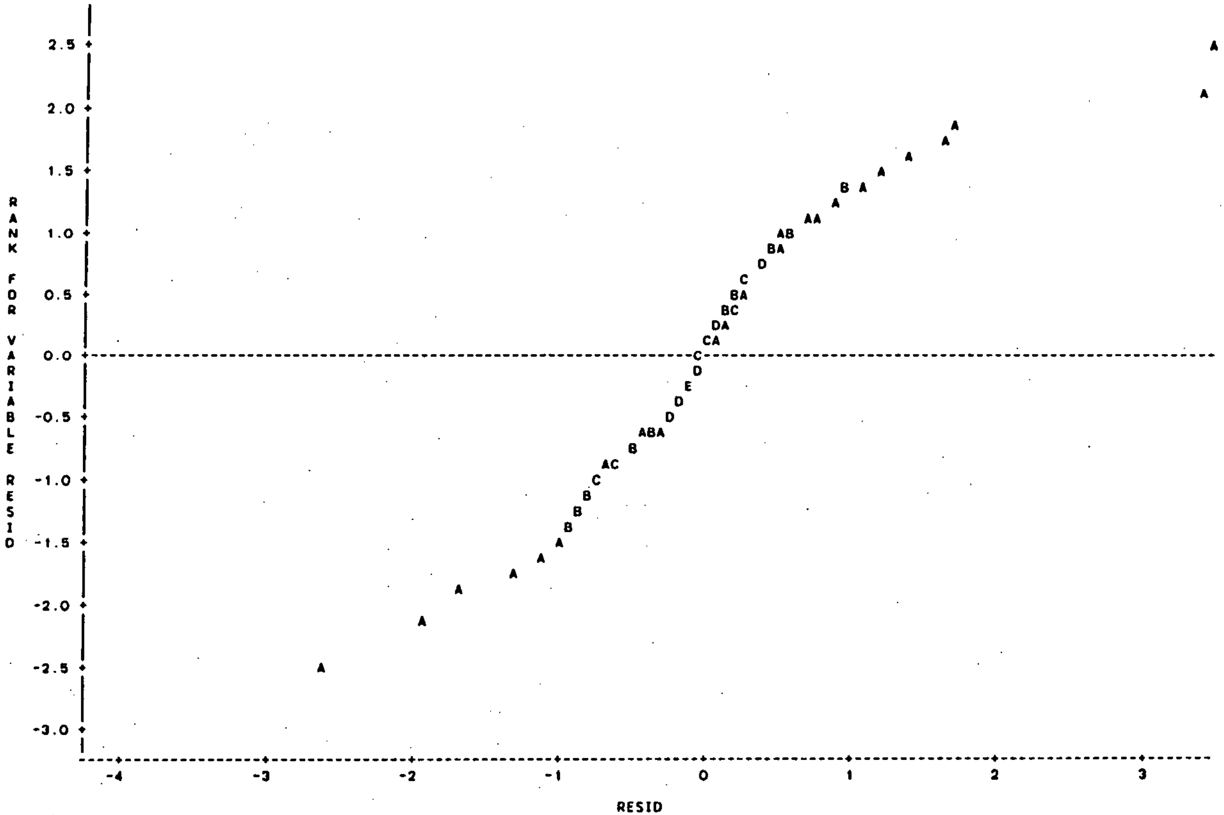
(1) 1978

(2) 1979

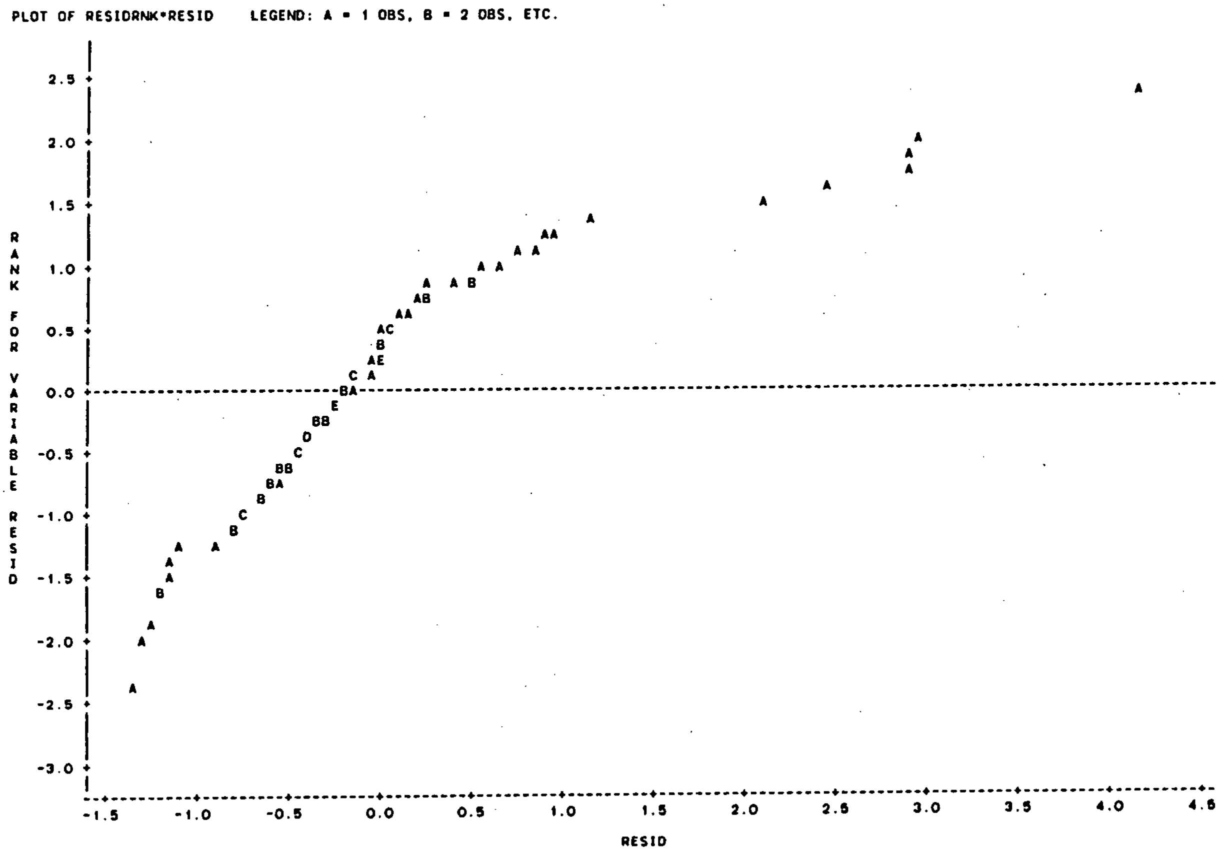
Figure 3.8: Normal plots for each year separately at station 020b.


PLOT OF RESIDRNK*RESID LEGEND: A = 1 OBS, B = 2 OBS, ETC.
RANK FOR VARIABLE RESID
2.5
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.5
-3.0
-2.5
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
RESID

(3) 1980


RANK FOR VARIABLE RESID
2.5
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.5
-3.0
-4
-3
-2
-1
0
1
2
3
RESID

(4) 1981


PLOT OF RESIDRNK*RESID LEGEND: A = 1 OBS, B = 2 OBS, ETC.
RANK FOR VARIABLE RESID
2.5
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.5
-3.0
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
RESID

(5) 1982