Data Analysis in Real Life Environment:
Ins and Outs of Solving Problems
J.-F. Marcotorchino, J.-M. Proth, and J. Janssen (Eds.)

ACID PRECIPITATION, EFFECTS ON SMALL LAKE CHEMISTRY: AN ATTEMPT TO ANALYSE DATA PRESENTED BY E. MOHN AND R. VOLDEN

P. NAEVE, D. TRENKLER, and P. WOLF
University of Bielefeld, Federal Republic of Germany

Data provided by Mohn and Volden are analyzed by some exploratory methods. Although they are "quick and dirty" some definite conclusions can be drawn. For example there are differences in the water chemistry of the north and of the south of Norway. The importance of Bross' principle of primacy with respect to this data set is emphasized.

The proposers of this problem, together with the data, suggest to go along the following lines:

1. Comparison of the water chemistry in the south of Norway with the north of Norway.

2. Are there any relations between the water chemistry and the precipitation?

3. Are there any relations between the changes in the water chemistry and the precipitation?

4. Is it possible to say anything about the development in time of the water chemistry during the period 1976 - 1981?

Expecially the last question shows, in connection with the ever growing world wide concern about ecological effects of man-made industrial civilisation, that statisticians do have a high responsability. Therefore, it seems mandatory in the beginning to give an impression of our approaches.

Some thoughts on statistical methodology

Surely, a classically trained statistician cannot overcome the temptation to apply some kind of t-test. For instance, he might calculate the differences in the water chemistry for the longest time span (i.e. pH (1981) - pH (1976), etc.) and test the hypothesis: no change in means, i.e. zero mean. He will end up with the following table:

variable	t-test	p two sided
pH	1.113	$0.5 > p > 0.2$
[SO4]	-1.7356	$0.1 > p > 0.05$
[NO3]	-1.6794	$0.2 > p > 0.1$
[Ca]	-1.2036	$0.5 > p > 0.2$

Table 1: Results of t-test

Be sure, he himself will guard his findings against objections concerning the assumptions necessary to apply the t-test. A plot like the following

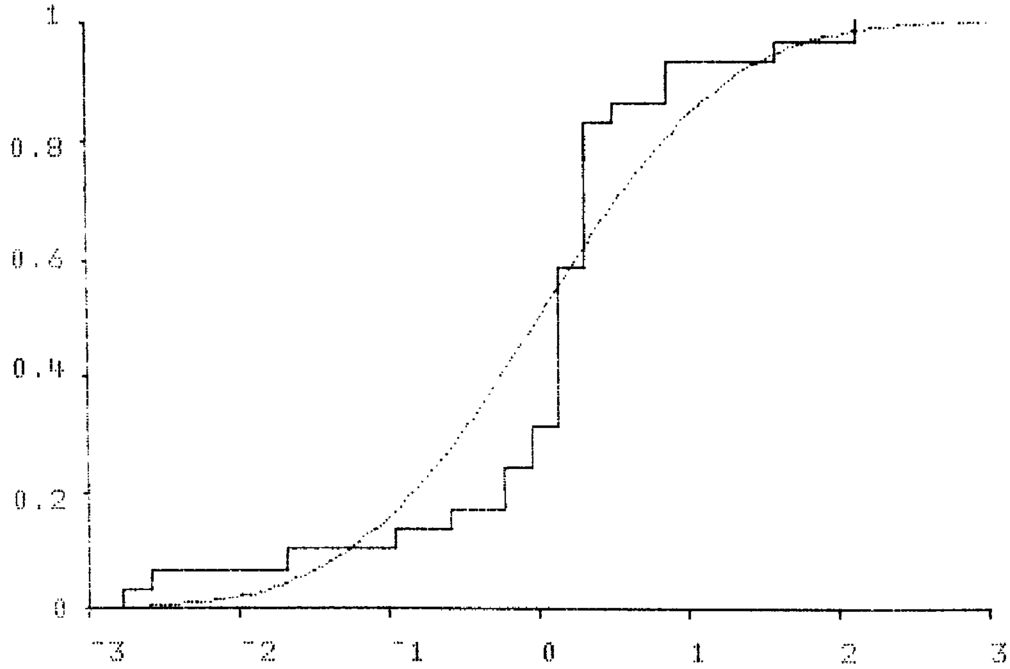

Figure 1: Empirical distribution of standardized NO3-differences 1981-1976

or a Lilliefors test for normality cast some doubt on his line of approach.

Either he might apply some kind of transformation e.g. square root, or logarithmic or Box - Cox, or move to the nonparametric branch of statistics. But transformations might lead to difficulties when one tries to interprete results, e.g. what would be the chemical explanation for a high correlation between SQRT([SO4]) and LOG([Ca])? Thus, our statistician devotes himself to the nonparametric approach and, choosing the Wilcoxon test, ends up with the following table:

variable	value	p two sided
pH	1.129	~ 0.26
[SO4]	-1.4	~ 0.16
[NO3]	-1.9	~ 0.06
[Ca]	-1.81	~ 0.07

Table 2: Results of Wilcoxon - test

But even now, one last hurdle has to be mastered: What is an appropiate α - value (probability of an error of the first kind) to wire statistician's most beloved message: "something is significant"? Usual values like α = 0.05, 0.01 or 0.005 in this investigation often suppressed that message. As a statistician, one can be satisfied with this finding, but as a concerned citizen, one remains unsatisfied.

There are two ways out. Firstly, one might advocate what Bross calls the principle of primacy. "... When we are dealing with a public health issue involving possibly hazardous technology, the benefit of the doubt must go to the public and not to the technology..."
This could mean, in this context, that p - values up to perhaps 0.2 should be sufficient to set the mechanism of that principle to work, i.e. ask the authority for actions. Secondly, one might recall John Tukey's saying: "Look at your data!" Usually, this is not done when one applies tests. Tests have a strong tendency to end up as canned computer programs which are applied automatically. Someone throws the data in and the statistician only looks at the results - if the program did not make him obsolete by printing a nice message like: "This is significant at the $\alpha = 0.05$ level." It is just one more step to GIGO, i.e. garbage in, garbage out. But Ehrenberg points out: "The need is to let the data speak." From here on, it is a logical step to advocate exploratory data analysis. Exploratory analysis is an ever growing body of helpful technics for looking at someone's data and letting them speak. But it is more than just a bundle of useful technics and procedures, it is an attitude as Tukey puts it. Exploratory analysis, in the long run, cannot be done with fixed tools and routine lines of approach. Although, at a first glance, the non - routine data analytic applications use many of the standard procedures and data processing tools, there is a distinguishing characteristic - the necessity to allow for human thought and intervention, as Heiberger puts it.
This was the spirit that guided us in our analysis. We took our supplies out of the arsenals of confirmatory data analysis i.e. all kind of tests as well as out of exploratory data analysis e.g. box - plots, stem and leaf, Q-Q-plots etc.. So, well equipped we went out on the hunt for patterns, trends and tendencies, gracefully collecting significant results in the old fashioned way, but also satisfied if only the principle of primacy could be claimed for our findings.

The following two figures show some typical "playing" with the data:

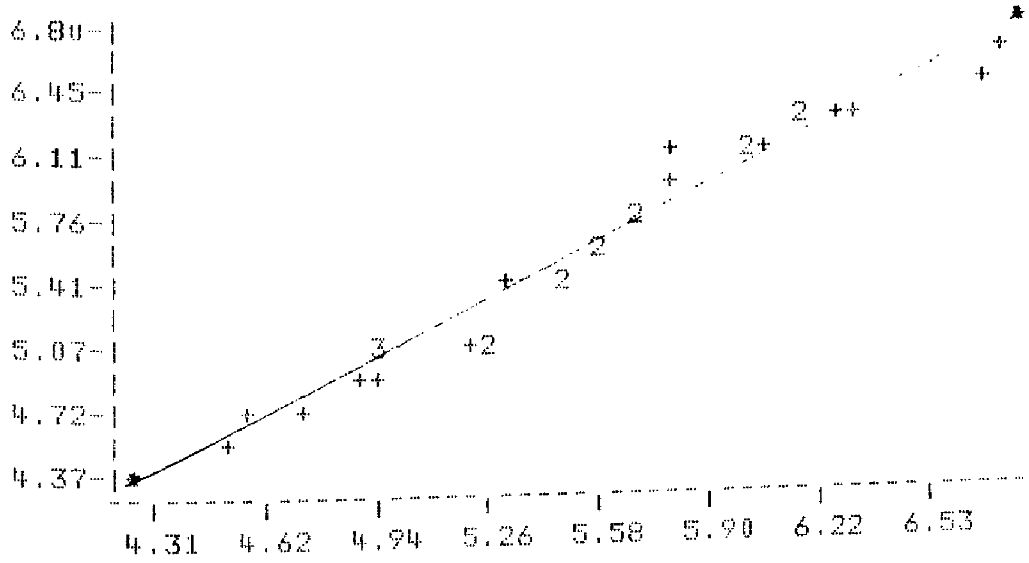

Figure 2: Q-Q-Plot of pH 81 against pH 76

To make it is easier to grasp deviations from a horizontal line, the plot was modified to:

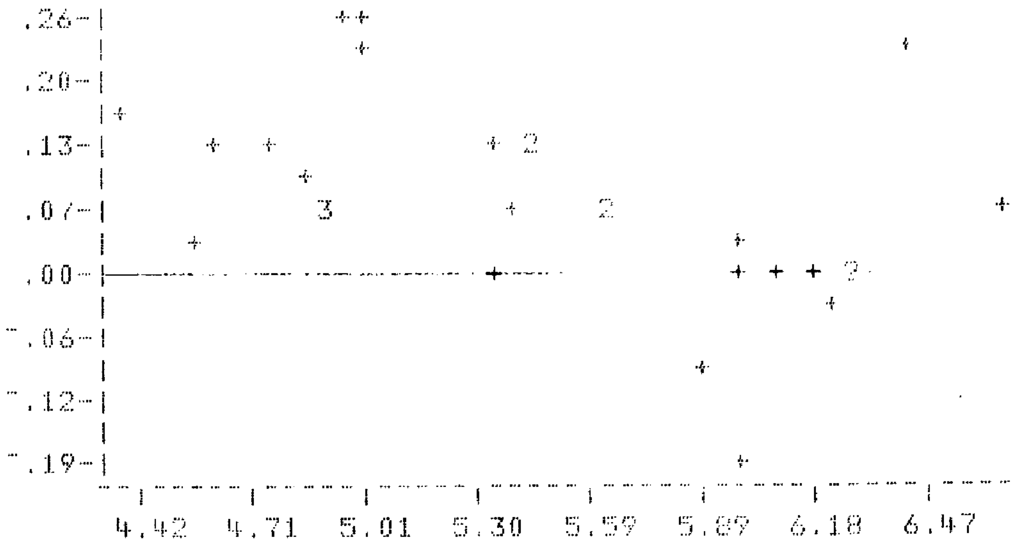

Figure 3: Modified Q-Q-Plot values as above

Both plots might be interpreted as follows: if both samples are from the same population, the points should lie on the 45^o-line (Figure 2) or on the horizontal line through $\Delta pH = 0$. (Figure 3). Sample fluctuations might be taken into consideration by testing for significant deviation from these lines. But even if this test fails, one can clearly see a pattern. Most points lie above these lines, i.e. there is a tendency of increasing pH-values.

Data base

Unfortunately, the data presented by Mohn and Volden contain many missing values. For our studies we extracted two data sets.

Data set 1: The data for the water chemistry in 1976
Data set 2: All data from the years 1976, 1977, 1978,
1981 for water chemistry and precipitation
for those lakes with no missing values (n=29)

We thought that by this selection criterion, we could keep the uncontrolled source of variation (for instance, there are other ions influencing the pH-values which are not included) within our data as homogeneous as possible. We enlarged data set 2 by a pair of coordinates for each lake which we took from a rather crude coordinate system. These coordinates enabled us to make a somewhat finer analysis of geographic regions than with the proposed classification of north - south lakes by the lake number.

Comparison of the water chemistry in the south of Norway with the north of Norway

In answering this question there are two points to be discussed: Firstly, we try to motivate a partition of the whole Norwegian area. From this we, secondly, tackle the above question.

Why look at a partition of Norway?
At the first glimpse we were frustated by the complexity of the data because they revealed no obvious structure, although we had supposed some to exist inherently. One of the main goals of data analysis is to display hidden structures and group the data. A geographical partition seems worthwhile, since the water quality of the lakes depends on the water quality of the precipitation and this, in turn on geographical parameters - even though the exact mechanisms are not well understood as yet. In this way a suitable partition leads to stronger results, hopefully anyway.

Our search for a partition:
We considered those lakes for which all data were completely given. So, we examined 29 lakes, the majority of which is located in the southern half of Norway. To characterize the position of the lakes, we defined individual x - and y - coordinates. Surprisingly we obtained some patterns by the technics to be described in the sequel:

a) As an example for the first technic, the NO3 values in 1976 were arranged in an ascending order and accordingly linked in a "map". (See Figure 5). Doing this, we obtained a map for each of the variables under consideration.

b) To get a better impression of local pollution, we plotted the observations of the variables against the respective lake coordinates. Explicitly we obtained plots that characterize four directions by defining the abscissa as follows:

construction of abscissa by coordinates	direction
x	W → E
y	S → N
x + y	SW → NE
x - y	NW → SE

Table 3: Coordinates and directions

Have a look at Figure 4.

c) Subsequent box-plots confirmed the found structure.

Some observations:

a) Low pH-values (for example pH-values less than the dangerous limit of 5) are only located in the south of Norway.

b) The highest concentration of SO4-values (lake water) can be observed in lakes situated in the south-east. (Figure 4).

c) Since there is a higher Ca-concentration in the east, the pH-values are not a sufficient indicator for the degree of pollution.

d) A more appropiate one may be NO3. Once again the lakes in the south yield the higher NO3-values. (Figure 5).

In summarizing these results, we can say that there are local differences. Such differences appear to be most significant along the direction north, middle-west, south-east. We feel that with respect to all kinds of ecological problems, the south-east deserves the analyst's special attention!

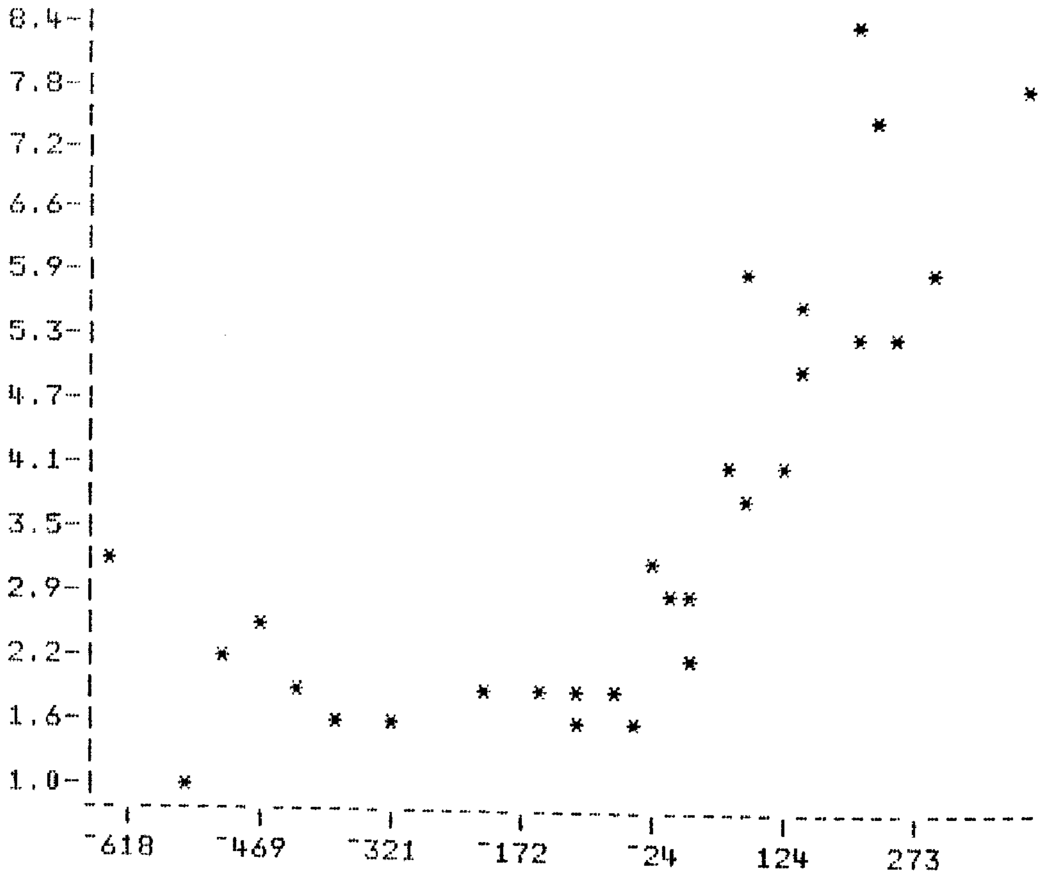

Figure 4: SO4 and direction NW → SE

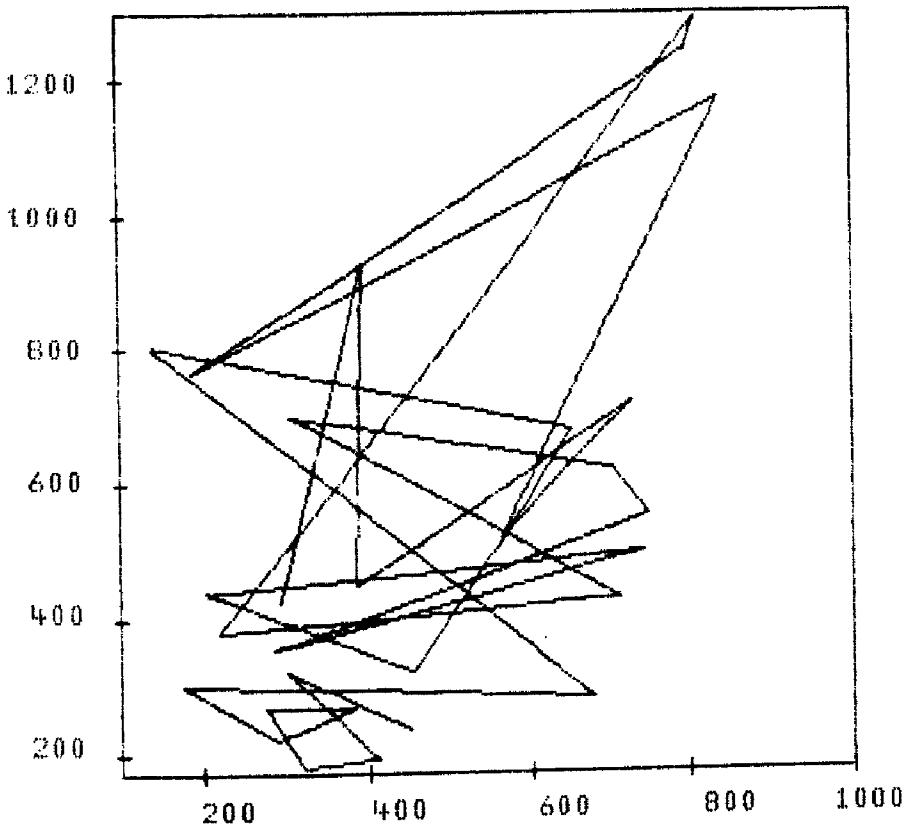

Figure 5: A map of NO3 in 1976

In order to compare the north with the south, only the year 1976 provides sufficient data. Consequently our conclusions mainly relate to that year. Upon referring to the before mentioned partition of Norway, we now distinguish between S (lakes in the south-east), M (lakes in the middle, i.e. the middle-west) and N (lakes in the north of Norway). To give an impression of the inherent structure of the respective data, we display a comparative stem-and-leaf plot for the pH-values in S, M and N.

It reveals the fact that the direction S → M → N delivers an increase in pH-values, which confirms the presumption that the progress of acidification of lakes in the north is not as extreme as for those in the south. Interestingly enough, things look quite similar for the precipitation H+ - concentration. Taking into account that there were data for only eight N - lakes available, the comparative stem-and-leaf plots show a decrease of H+ - concentration.

```
Factor of Stem:  1
1∘2 means 1.2

    4∘        |      4∘          |      4∘
  1 4∘3       |      4∘          |      4∘
  4 4∘567     |      4∘          |      4∘
(3)4∘999      |    2 4∘89        |      4∘
  5 5∘00      |      5∘          |      5∘
  3 5∘44      |    5 5∘334       |      5∘
    5∘        |    7 5∘66        |    4 5∘6677
    5∘        |    8 5∘9         |      5∘
  1 6∘0       |  (5)6∘00122      |    8 6∘0001
    6∘        |    4 6∘334       |   15 6∘3344444
    6∘        |    1 6∘7         | ( 4)6∘5577
    6∘        |      6∘          |   16 6∘888999999
    7∘        |      7∘          |    7 7∘0012
    7∘        |      7∘          |    3 7∘3
    7∘        |      7∘          |    2 7∘55
    7∘        |      7∘          |      7∘
```

Figure 6: Stem-and-leaf of Ph in S, M, N

Although not explicitly shown here, similar results are obtained for the other precipitation-values.

```
Factor of Stem:  10
1∘2 means 12

                 1∘     |      1∘
                 1∘     |    3 1∘233
                 1∘     |      1∘
                 1∘     |    5 1∘67
                 1∘     |    7 1∘88
                 2∘     |    8 2∘0
                 2∘     |  (1)2∘2
                 2∘     |    8 2∘5
                 2∘     |    7 2∘67
               1 2∘9    |    5 2∘89
                 3∘     |    3 3∘0
               3 3∘23   |      3∘
               4 3∘4    |    2 3∘4
                 3∘     |      3∘
                 3∘     |      3∘
               5 4∘0    |    1 4∘0
             (2)4∘23    |      4∘
               5 4∘5    |      4∘
               4 4∘6    |      4∘
               3 4∘8    |      4∘
               2 5∘1    |      5∘
               1 5∘2    |      5∘
                 5∘     |      5∘
                 5∘     |      5∘
                 5∘     |      5∘
```

Figure 7: Stem-and-leaf of H-prec. in S, M

The inspection of notched box-plots provides additional evidence for the statement made above.

Note that in Fig. 7 and 9 the north is omitted due to lack of data.

```
 4.44  4.79  5.14  5.50  5.85  6.20  6.56  6.91  7.26
--+-----+-----+-----+-----+-----+-----+-----+-----+----

      [---+----]Φ
*-----Φ   +     Φ---*        ⊕              (South)
      Φ---+-----Φ

                   Φ---[----+-----]
         *---------Φ        +     Φ------*   (Middle)
                   Φ--------+-----Φ

                                   Φ-[--+-]---Φ
        (North)       oo*----------Φ    +     Φ---------*
                      2            Φ----+-----Φ

--+-----+-----+-----+-----+-----+-----+-----+-----+----
 4.44  4.79  5.14  5.50  5.85  6.20  6.56  6.91  7.26
```

Figure 8: Notched box-plots of pH in S, M, N

```
13.89 18.29 22.69 27.10 31.50 35.90 40.31 44.71 49.11
--+-----+-----+-----+-----+-----+-----+-----+-----+-----
                              Φ---[-------+-----Φ ]
      (South)          *------Φ           +     Φ------*
                              Φ-----------+-----Φ

        [-----+----]--Φ
 *------Φ     +       Φ-------*       o     (Middle)
        Φ-----+-------Φ

--+-----+-----+-----+-----+-----+-----+-----+-----+-----
13.89 18.29 22.69 27.10 31.50 35.90 40.31 44.71 49.11
```

Figure 9: Notched box-plots of H+ - prec. in S, M, N

We next administered several statistical methods to see if the observed differences were significant. Among those were the parametric t-test, the Mann-Whitney-Wilcoxon-test, the Smirnov-test and the Kruskal-Wallis-test. We believe that the prevalence of distribution-free methods is sufficiently justified, since the assumption of normal-distributed data seems at least doubtful.

Let us have a look at the respective empirical distribution-functions of the pH-values:

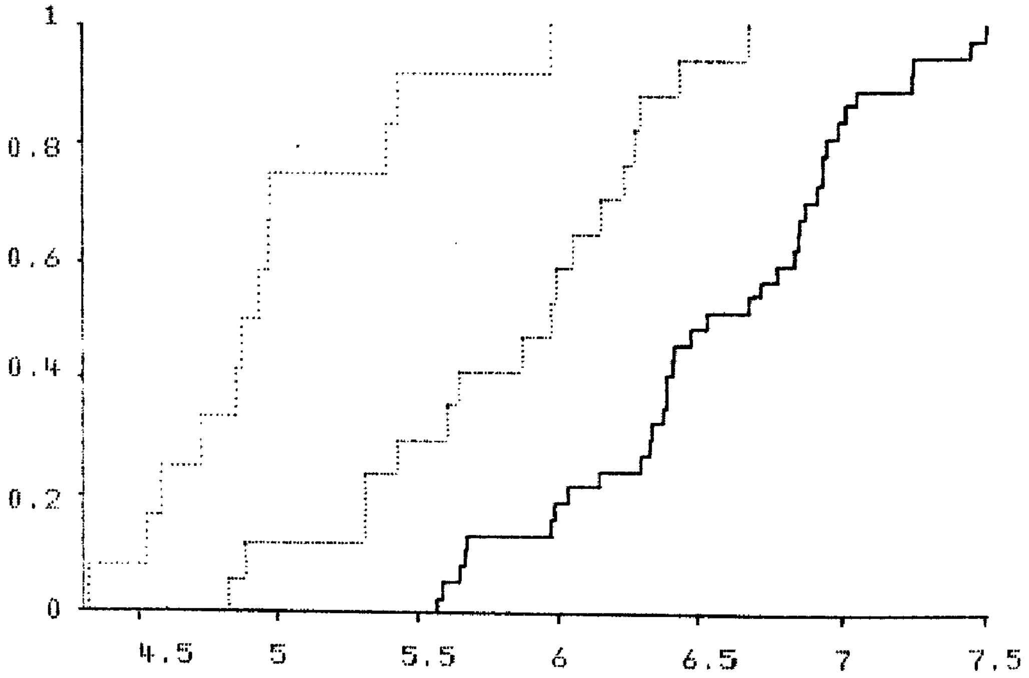

Figure 10: Empirical Distribution of pH-values of S-, M- and N-lakes

Once again, we are encouraged to formulate the null hypothesis:

H0: The distribution of pH-values is identical for south, middle and north.

A distribution-free test for this kind of hypothesis is the Kruskal-Wallis-test. It rejects H0 at a critical level $\alpha < 0.001$. To see which pairs of distributions differ significantly, Conover discusses a method which is applicable if and only if the null hypothesis is rejected. At a level of significance $\alpha = 0.05$, we may say that all pairs of populations differ. By the same token, we are led to say that for the other populations (S04-, N03-, Ca-concentration) similar results follow. The difference between S and N is always significant. These findings are confirmed by other tests. For example, comparing N-pH- and S-pH-values, the Smirnov-test rejects H0 at a critical level $\alpha < 0.01$.

What may we conclude on the basis of such findings?
In 1976 there were considerable differences in the water chemistry of lakes in the south and in the north of Norway. Furthermore, this also seems to be the case concerning the H+ - and S04 - concentrations in the precipitation.

This latter point has to be observed with care, since only a small amount of data was available to make this conjecture more profound. Assuming that the acidification process for the lakes in the north has been as inert over the past few years as that, which we observed for the lakes in the south, we arrive at the conclusion that things in the north do not look as bad as in the south of Norway.

Are there any relations between the water chemistry and the precipitation?

To summarize our findings the answer is affirmative. We will demonstrate these facts just by showing two series of plots standing for a multitude of similar statistical analyses.

Our first plot is a scatterdiagram where all pairs ([S04]-lake, S04-deposite), for the years 1976-1978, are included. This plot does not exhibit any clear-cut pattern:

```
9.6-|                *
    |               *      *                         *
8.4-|                *
    |                 *   *                           *
7.2-|          *   *
    |          *                        *
6.0-|     *  2 2 *    *   * *
    |     ***  * *            *
4.8-|                        **
    |        *  *  * **     *
3.6-|                     *        *
    |*    * *** * * * * *  *
2.4-|* *34*  **** **      * *       **             *
    |2 **32 2  * *   * 2
1.2-|  *                               *
    --|-----|-----|-----|-----|-----|-----|-----|------
     403   681   960   1238  1516  1795  2073  2352
```

Figure 11: Scatterdiagram [S04]-lake, S04 - deposite for 1976-1978

Our next idea was to do the same kind of plot for [S04]-lake against [S04]-rain. Figure 12 gives the impression of some kind of (linear) relation.

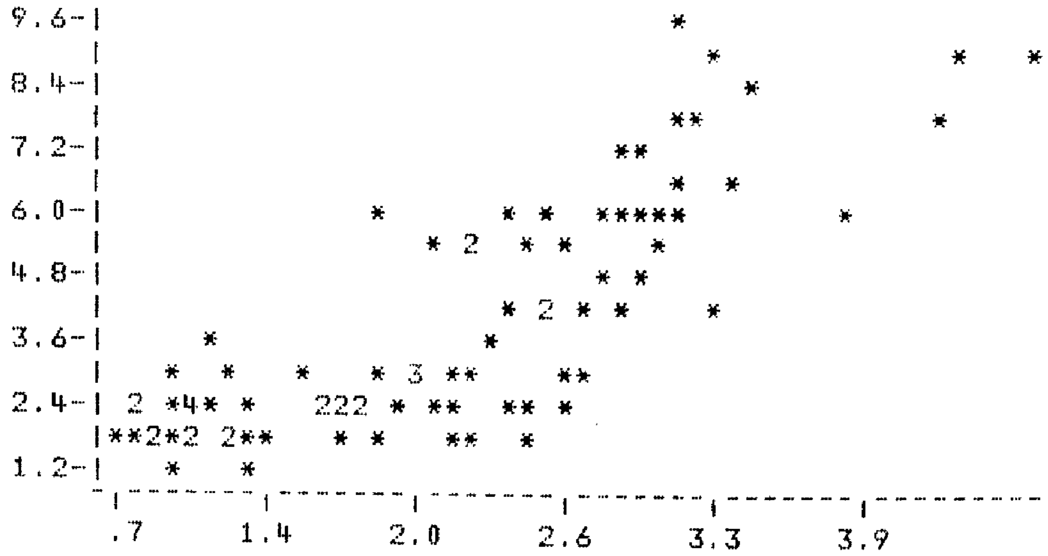

Figure 12: Scatterdiagram [S04]-lake, [S04]-rain for 1976-1978

The relationship can be seen more clearly when one plots the (unweighted) mean of [S04]-lake over these three years against the corresponding mean of [S04]-rain:

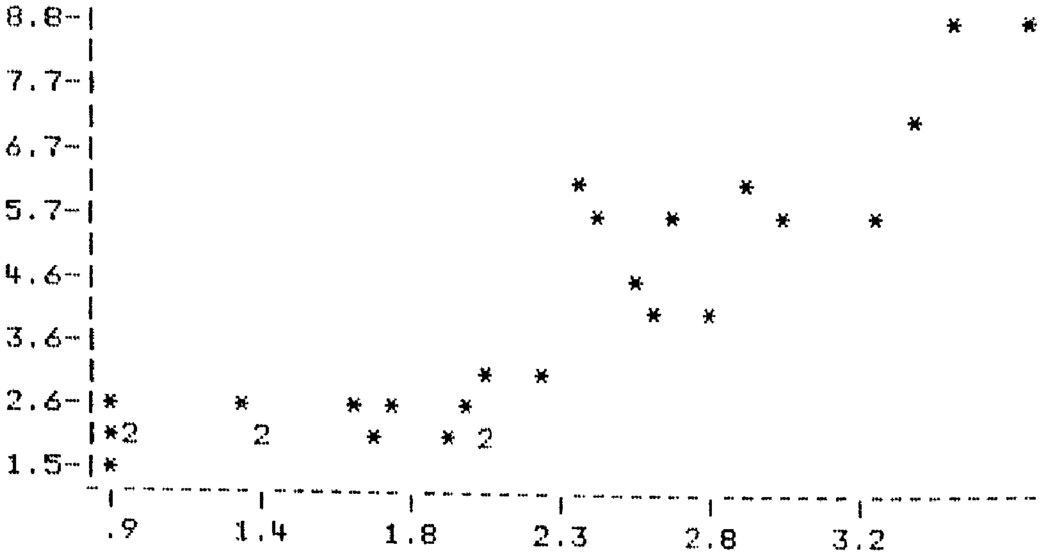

Figure 13: Scatterdiagram mean [S04]-lake, mean [S04]-rain

Secondly we searched for a relationship between pH-lake and pH-rain. For [Ca] should have a certain buffer effect we plotted scatterdiagrams like that of figure 14:

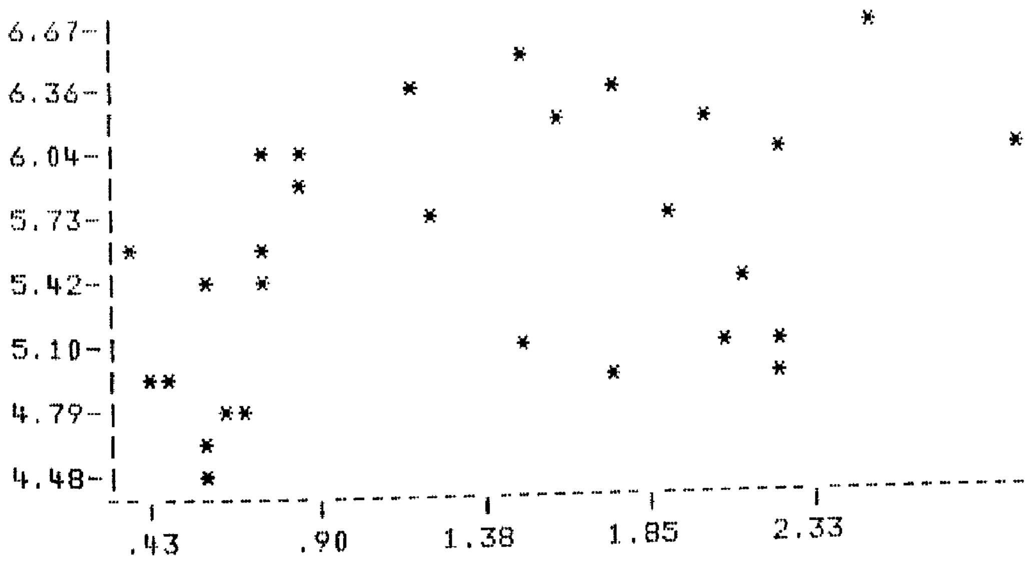

Figure 14: Scatterdiagram of pH-lake against [Ca] for 1976

There seems to be no convincing pattern. The data for the other years give similar results. Next we investigate the variables pH-lake and [SO4]-rain:

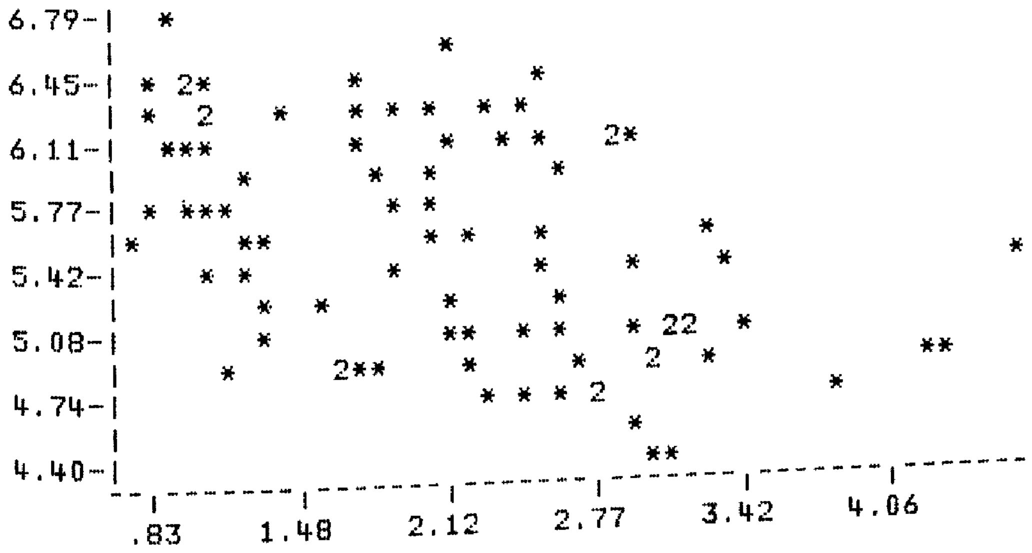

Figure 15: Scatterdiagram of pH-lake against [SO4]-rain for 1976-1978

Again no pattern is visible. Then we developed the idea that it might be necessary to look at more than two variables at a time. So we thought of weighting the effect [S04]-rain should have inverse proportional to the [Ca] of the lake.

The next two plots present the results for the year 1976 and for (unweighted) mean pH against (unweighted) mean [S04] ÷ [Ca]. The mean was taken over the years 1976-1978:

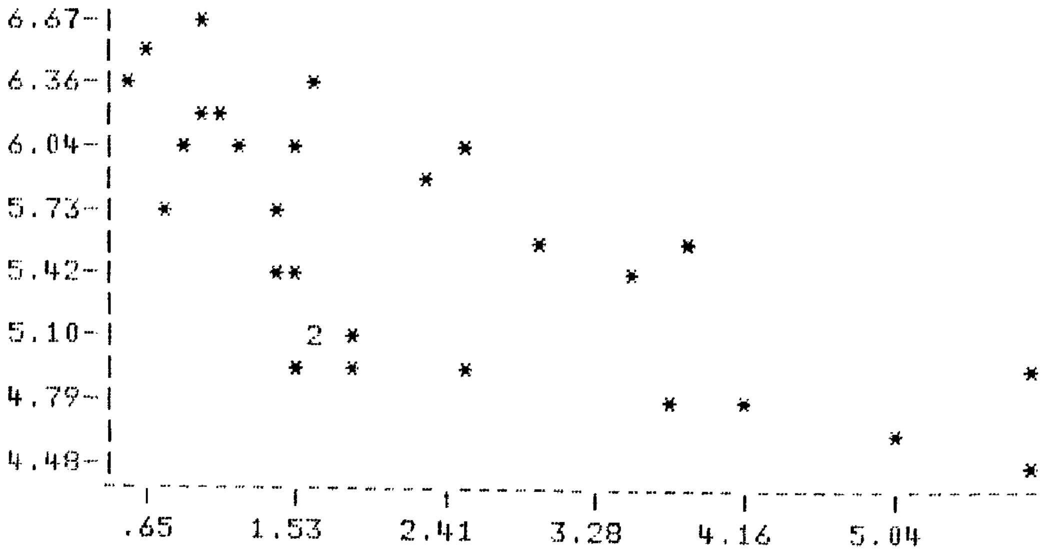

Figure 16: Scatterdiagram pH-lake against [S04]-rain : [Ca] for 1976

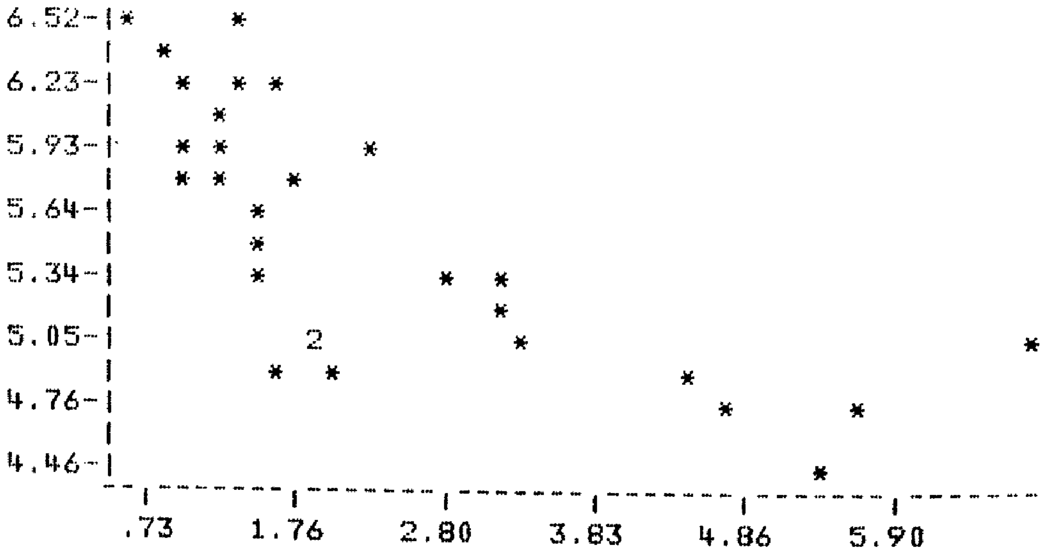

Figure 17: Scatterdiagram mean pH-lake against mean [S04]-rain ÷ mean [Ca]

We think the pattern is clear cut and convincing. Certainly one should look for similar technics to include more than three variables.

Are there any relations between the changes in water chemistry and precipitation?

We were not able to find any pattern which would cast some light on this question. But if one considers the results concerning question 2, some kind of relation should be present.

We think the failure to detect that relation is mainly due to two effects. Firstly, the data base is rather limited with respect to the number of lakes included, as well as to the time interval covered by the data. Secondly, there might be high variability within the water chemistry data due to important variables not included, such as amount of snow, concentration of ions in the melting water, etc. This high variability might disturb the relational pattern within the data at hand.

Is it possible to say anything about the development in time of the water chemistry during the period 1976-1981?

Let us look at the data from a pessimistic point of view. Is it realistic to say that things have become worse? The following comparative stem-and-leaf-plot is inconclusive.

```
Factor of Stem:  1           1∘2 means 1.2

    4∘         |    4∘          |    4∘           |    4∘
  1 4∘3        |  1 4∘2         |    4∘           |    4∘
  2 4∘5        |    4∘          |  1 4∘4          |  2 4∘55
  4 4∘67       |  5 4∘6777      |  3 4∘66         |  3 4∘7
  9 4∘89999    | 10 4∘88889     |  9 4∘889999     |  8 4∘89999
 11 5∘00       | 12 5∘01        | 11 5∘01         |    5∘
 13 5∘33       | 14 5∘23        | 14 5∘233        | 12 5∘2223
( 3)5∘444      |( 1)5∘4         |( 1)5∘5          |( 3)5∘455
 13 5∘66       | 14 5∘7         | 14 5∘6667       | 14 5∘677
 11 5∘9        | 13 5∘88        | 10 5∘889        | 11 5∘88
 10 6∘0001     | 11 6∘00011     |  7 6∘01         |  9 6∘0001
  6 6∘2233     |  6 6∘233       |  5 6∘223        |  5 6∘233
  2 6∘4        |  3 6∘44        |  2 6∘45         |    6∘
  1 6∘7        |    6∘          |    6∘           |  2 6∘77
    6∘         |  1 6∘8         |    6∘           |    6∘
```

Figure 18: Stem-and-leaf of pH-lakes for 76-81

For example, the medians do not considerably move upward over the years. We found no major tendency for the other variables either. To extract a trend is obscured by the fact that some years seem to yield extreme conditions. The following notched box-plots may illustrate this point.

```
 20   61  102  142  183  224  264  305  346  386  427  468  509  549
-+----+----+----+----+----+----+----+----+----+----+----+----+----+----

  [Φ--+---]-----Φ
*---Φ +         Φ----*            ●  ●       ●        (1976)
    Φ-+---------Φ                 2

      Φ---[---+-]Φ
 *----Φ       +  Φ----*                    ●   ● ●    (1977)         ●  ●
      Φ-------+--Φ

 Φ[-----+----]-----Φ
*Φ      +          Φ---------------*     o            (1978)
 Φ-----+-----------Φ

    [-+--]----Φ
*---Φ +       Φ--------* o                  ●         (1981)
    Φ-+-------Φ

-+----+----+----+----+----+----+----+----+----+----+----+----+----+----
 20   61  102  142  183  224  264  305  346  386  427  468  509  549
```

Figure 19: Notched box-plots of NO3-lakes for 76-81

The year 1977 acts out of character. But note that, since the overall amount of precipitation for that year was very low, explanations are even more difficult.

On the other hand, maybe we should take into account that we are confronted with matched observations and that annual developments may be reflected by differences. In addition, as noted earlier, a local partition should be taken into consideration. These aspects will be found in the notched box-plots on the next page (figure 20).

This, surely, is a kind of magnifying glass, but from our point of view the complexity has increased. Sometimes there are local differences in the annual differences, sometimes not. One may have the impression that SO4-values of S decrease while those of M increase (look at 81-77 and 81-76).

In all, we still found no general tendency. A statistician might not be frustrated by this result, but this does not mean that we are content as citizens. On the contrary. Since we hold that human beings are strongly linked with nature we are worried about the state of affairs.

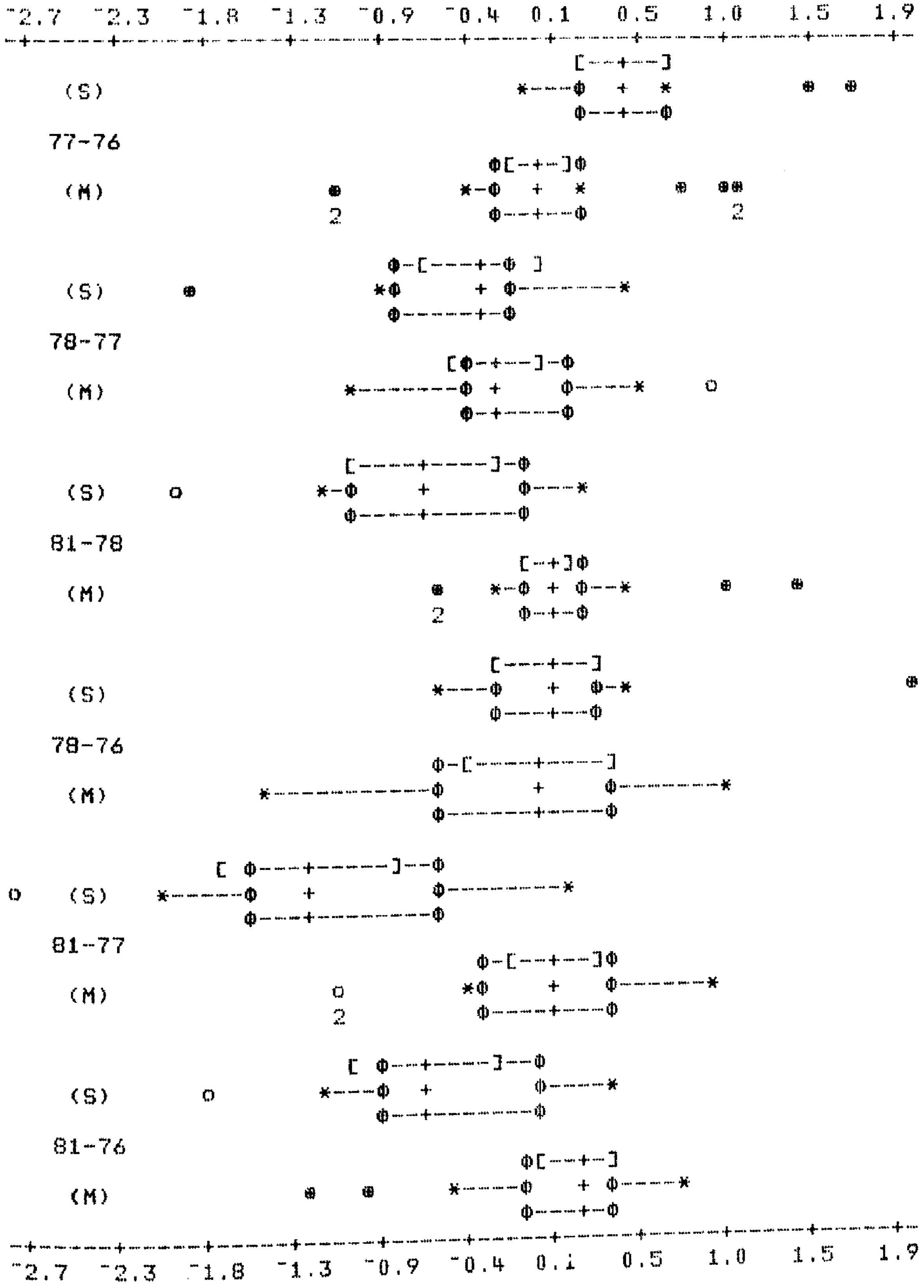

Figure 20: Notched box-plots of annual S04-differences for the lakes in S and M

Concluding remarks

We were able to answer the first two questions positively. There certainly is a (significant) difference in the water chemistry of the north and of the south of Norway. An even finer structure can be isolated. And there is a relationship between the absolute values of water chemistry and precipitation. No definite evidence was found for a connection between changes in water chemistry and precipitation. The analysis of the development in time did not bring very clear results. It should be pointed out again that the time span covered by the data is rather short. Continuous measurements are called for.

For the long-run development of a lake, one might resort to a kind of ergodic argument. If we take it for granted that in a preindustrial age there was not much difference among all Norwegian lakes, one might choose the acid profile of lakes along a north - south - line as equivalent to the time development of an individual lake.

Last but not least, we think our considerations demonstrate that explorative data analysis is a valuable tool. Although the data base was not very good some results could be found. As a final remark, we once more would like to turn attention to the principle of primacy formulated by Bross.

References

Bross, I.D.J. Scientific strategies to save your life.
Dekker, New York 1981

Conover, W.J. Practical nonparametric statistics 2.ed.
J. Wiley, New York 1980

Ehrenberg, A.S.C. Data reduction.
J. Wiley, London 1975

Heiberger, R.M. Software for statistical theory and practice.
Technical report 42, Dept. of Statistics.
The Wharton School, Univ. of Pennsylvania

Tukey, J.W. Exploratory data analysis.
Addison - Wesley, Reading 1977

Tukey, J.W. We need both exploratory and confirmatory.
The American Statistician, vol. 34