ASSESSMENT OF THE QUALITY FOR THE NADP/NTN DATA BASED ON THEIR PREDICTABILITY

By

SAULATI KOKU KOMUNGOMA

B.Sc., University of Dar-es-Salaam, 1980

A THESIS SUBMITED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

THE FACULTY OF GRADUATE STUDIES

DEPARTMENT OF STATISTICS

We accept this thesis as conforming

to the required standard

....

....

THE UNIVERSITY OF BRITISH COLUMBIA

April 1992

In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

(Signature)

Department of Statistics

The University of British Columbia
Vancouver, Canada

Date 29/4/1992

DE-6 (2/88)

Abstract

Three methods are used to predict the ion concentrations of a particular station using the concentrations of the other stations, for the data produced by the National Acidic Deposition (NADP) Network/ National Trends Network (NTN), during the period of 1983-86. We relate the degree of predictability to the quality of the data. Stations are ranked in the order in which they would be dropped if the network were, hypothetically, to be reduced in size. The agreement of the ranks given by different methods is assessed.

Our study uses monthly volume weighted mean concentrations for each of the three selected ions, investigated one at a time. Since there a large number of stations (86 for hydrogen, 81 for each of the remaining ions) and only 48 months, analyses was carried out on clusters of stations. It was not possible to perform an ordinary regression analysis with a lot of missing data, so the analysis is done with missing values replaced by their estimates.

Contents

List of Tables

List of Figures

Ackowledgement

I wish to thank members of the Department of Statiatics and Faculty of Graduate Studies at the University of British Columbia for their continued assistence throughout my studies. In particular I wish to thank my thesis supervisors, Professor J. Zidek and Profesor N. Heckman for their help and continued support during my research and thesis write-up. My special thanks goes to Professor M. Stone who suggested the idea of cross-validatory assessment in network design to my supervisors Professor J.V. Zidek.

I would also like to thank my sponsors, the African Women 2000 Award who supported me financially.

It is my pleasure to acknowledge the kindness and assistence extended to me by my husband Mr Ayub Komungoma. I do not forget others who gave me moral support.

CHAPTER 1

BACKGROUND

1.1 Introduction

The National Deposition/National Trend Network (NADP/NTN) is one of the networks which collect rainfall chemistry data in different locations in the U.S. Each location is called a station and is identified by a station code. Details of the network and the data are given in the next section. The goal of this study is to assess how well the rainfall chemistry of a particular station can be predicted from chemistries of other stations. This information might be used to reduce the size of the network, if the need arises, by dropping from the network a station whose rainfall chemistry is satisfactorily predicted from other stations' rainfall chemistries. The rainfall chemistries studied are the concentrations of 3 ions, namely hydrogen, sulfate and nitrate. In the following section, we discuss the nature of acidic deposition and its effect on biota.

Various methods of predicting one station's chemistry from others are considered, namely, ordinary regression, regression using a Bayesian approach and Stone's cross validatory procedure. Descriptions of each method are included in the next chapter. Each month's rainfall chemistry at a particular station is predicted from the rest using each method, one at a time. This is, in turn, used to get a prediction error for each month, one at a time, for that particular station. The average squared prediction error at each station is used to assess each station's predictability. The smaller its average squared prediction error, the easier the prediction of a station. Stations are ranked by the above criterion. In this way we can rank the stations in the order in which they might hypothetically be dropped from the network, if necessary. Finally the rankings from the three different methods plus the rankings of Wu and Zidek (1992) from an entropy based approach are compared. A brief review of this entropy approach is given

in Section 1.3. For more details, see Wu and Zidek (1992).

1.2 Effect of Surface Water Chemistry to Biota

The two main negatively charged ions that play a major role in the process of acidic deposition are sulfate and nitrate. These ions combine with hydrogen ions to form acidifying chemical compounds (sulfuric and/or nitric acid).

Acidic deposition causes surface water to lose its acid neutralizing capacity (ANC), which results in increased acidity (lower pH) and increased inorganic aluminum, which is toxic to aquatic organisms.

The extent to which acidic deposition causes surface water acidification is determined by the process occurring in the surrounding watershed. When water moves through the watershed, various processes change their chemical composition. The most prominent processes that take place are those that neutralize acids and release base cations (positively charged ions such as calcium and magnesium). One of these processes is mineral weathering, in which minerals gradually dissolve with the passage of time. The other is a reaction in which the ions are exchanged in the soil, that is, the acidic hydrogen ion entering the soil is absorbed in the soil, replacing absorbed base cations, which in turn are released to the water.

As most surface waters are well buffered, with pH values between 6.5 and 8.0, waters in which acid neutralizing and acid generating processes are nearly in balance are most likely to be affected by acidic deposition.

Acidic deposition on surface water increases sulfate. The trend in many areas is that sulfate concentration increases as the rate of acidic deposition increases. However the nitrates remain low; although nitrate is a very important compound in acidic deposition, most watersheds retain it efficiently because of its importance in plant nutriency. On the other hand as acid inputs to a watershed increase, there is a nearly universal response

of increase in magnitude of acid neutralizing reactions that produce base cations. In watersheds almost all of the acid input is neutralized, with no change in pH or ANC of the surface waters. Even in the most sensitive waters, a substantial fraction of the acid input is neutralized by the processes that release base cations. As ANC and pH decrease, aluminum increases. When pH declines, aluminum, which is found in nearly all soil minerals, is leached from the soils, causing concentrations levels in lakes and streams to rise. Dissolved organic carbons tend to decrease with increase in acid input. This process reduces the decline in ANC and pH, thus partially making up for increased acidity.

The harmful effects on aquatic life are not caused by ANC change alone, because aquatic organisms respond to many factors. Other factors affecting aquatic organisms are the change in pH and the release of calcium caused by acidification. The change in pH is the main variable that affects aquatic life.

Aluminum concentrations are always low in non-acidic waters. When the pH value decreases below 5.5, the concentration of aluminum increases, very often to toxic levels. Both a decrease in pH and an increase in aluminum can cause acidification toxic to fish and other biota.

In very dilute systems, low calcium levels could be stressful to fish, although in these waters the concentration of base cations increases in response to acidic deposition. Elevated base cations, especially calcium, may partially mitigate the toxic effects of low pH and high aluminum. Therefore, as surface water acidifies, the resulting combination of hydrogen ions, aluminum, and calcium determines the biological effects.

Some types of organisms are sensitive to the chemical changes that accompany acidification and thus can not grow, survive, or reproduce in acidified water. As acidity increases, these acid-sensitive species perish, resulting in the decline of species richness.

Phenomenon of surface water chemistry and its effect on biota have been of much concern to researchers. Data are collected all over the world to enable researchers to come up with definite conclusions about the response of aquatic life to acidification. Detail on acidic deposition is found in the National Acid Precipitation Assessment Program, 1990 Integrated Assessment Report.

1.3 Network and Data Description

The NADP/NTN is one of the networks in the U.S. which collect data on acidic deposition. This network collects weekly wet precipitation samples at more than 200 stations. These precipitation chemistry samples are analyzed by the Central Analytical Laboratory at the Illinois State Water Survey in Chicago, where ion concentrations are measured. Finally the data are transferred to the Acidic Deposition System (ADS) data base. This data base was established by the U.S. Environmental Protection agency at the Northwest Laboratory in the U.S. to provide an integrated centralized data base for the data collected by atmospheric deposition networks in North America. For more details about the NADP/NTN network and ADS refer to Olsen and Slavish (1986).

Our study uses monthly volume weighted mean ion concentrations rather than weekly means since there is a large number of weeks with no precipitation. The analysis is done on one ion at a time using stations with less than five missing monthly volume weighted means. As a result only 86 stations are used for the hydrogen ion analysis, and 81 stations in the analysis of the remaining ions. The data included in this study were collected between 1983 and 1986 inclusive, so a total of 48 monthly volume weighted means are used in the analysis. Because of the small number of monthly volume weighted means relative to the number of stations, it is necessary to restrict our analysis to small clusters of stations and proceed from cluster to cluster. The clusters given by Wu and Zidek (1992) are used. Clustering was done for each ion separately using the k-means

algorithm of Hartigan and Wong (1979). Given k, the number of clusters, the method seeks to find k clusters so that the within-cluster sums of squares are minimized. The method proposed by Krzanowski and Lai (1988) was used to select k, the number of clusters for each ion. The numbers, k, selected by the method in this study for different ions, range from three to six. The cluster sizes range from 2 to 47 stations. A logarithmic transformation of the data was performed prior to the analysis. In certain clusters with more than 20 stations the number of complete records over all stations is less than the number of stations in the cluster. For example, the number of complete records in one of the sulfate clusters with 36 stations is 19. So for the analyses presented in this thesis, the missing values are replaced by their estimates. More about the estimation of missing values appears in Section 1 of Chapter 2.

1.4 Review of the Entropy Approach

The purpose of an environmental monitoring network may be difficult to specify precisely. This presents a dilemma for the designer of such a network. Caselton and Zidek (1984) argue that the purposes of any network are in essence the reduction of uncertainty about some aspect of the world. They conclude that a rational design must minimize entropy, a measure of uncertainty.

The theory of entropy and its potential role in assessing the quality of the data generated by an existing network are described by Caselton, Kan and Zidek (1990). If X is an uncertain, i.e. random, quantity or vector of such quantities, and f is the probability density function of X, then the uncertainty about X is expressible by the entropy of X's distribution, i.e. by $H(X) = E[-logf(X)/h(X)]$, where, according to Jaynes (1963), h is a "measure" representing complete ignorance. The inclusion of h in this definition of entropy makes this index of uncertainty satisfy the natural requirement of being invariant under one-to-one transformations of the scale of X. Although the

uncertainty about X is regarded as being of primary interest, often its distribution is determined by the conditional density of X, $f(\cdot|\theta)$, given an unspecified parameter, θ, which is of interest in its own right. In this case the total uncertainty of (X, θ) must be indexed. Conditional on the available data, the total uncertainty is then indexed by the total entropy, defined by

$$H(X,\theta) = -E\Big[log[\frac{f(X,\theta|data)}{h(X,\theta)}]|data\Big] \tag{1}$$

where the expectation is taken over both X and θ, and

$$f(X,\theta|data) = f(X|\theta,data)\pi(\theta|data), \tag{2}$$

$\pi(\theta|data)$ being the posterior distribution of θ.

To assess the performance of stations in the network using the entropy approach, it is supposed that hypothetically, a specified number of stations, u, is to be dropped from the network and only g coordinates of X will be measured in future, with $u + g = p$, where p is the number of stations. After rearranging subscripts as necessary, let $X = (U, G)$ where U and G denote, respectively, the u and g dimensional vectors of values corresponding to the stations which are to be "ungauged" and "gauged". The process of measurement will eliminate the uncertainty about G assuming the measurement error to be negligible. The amount of uncertainty so eliminated would be $MEAS$ defined by $MEAS = -E[log[f(G|data)/h(G)]|data]$. The set of g stations would be chosen to maximize this entropy ($MEAS$). It can be shown that this same set of g stations can be found by minimizing $PRED + MODEL$, the residual uncertainty remaining after G is observed, where $PRED = -E[log[f(U|\theta,G,data)/h(U)]|data]$ and $MODEL = E[-log[f(\theta|G,data)/h(\theta)]|data]$.

The g stations that maximize $MEAS$ are considered to produce high quality data. On the other other hand, the u stations that maximize $PRED+MODEL$ (the residual uncertainty after G is observed) are regarded as producing low quality data.

Wu and Zidek (1992) apply this theory to assess the quality of the same data set as in our study. The analysis was done one ion at a time, since there are ion-to-ion differences in data quality. The stations were clustered and the entropy analysis done within each cluster. This was necessary because the number of stations is greater than the number of observations (48 volume weighted ion concentration means). If the size of the network were, hypothetically, to be reduced, then only the selected g stations would be retained.

In the implementation of the entropy theory, Caselton et al (1992) and Wu and Zidek (1992) found it was not computationally feasible to find the best subset of g gauged stations among all p stations in the cluster. Thus a stepwise suboptimization procedure was used. The first step consisted of finding $p-1$ stations which maximize $MEAS$. The station left out would be the first one to be dropped from the network, hypothetically speaking. The next step was to find the $p-2$ stations among the $p-1$ selected stations in the first step which maximize $MEAS$, and this yielded a second station which might hypothetically be terminated. This process continued until just one station was hypothetically left in the network. This exercise left the stations in ranked order, starting with the station having the lowest quality data and finishing with the station having the highest quality data.

Looking at the rank order within clusters, the station identified as having highest quality data is, in most cases, geographically isolated from the rest of the stations in the cluster. This seems reasonable since gauging such a station would be expected to substantially reduce the uncertainty. We may conclude that the entropy approach is promising as a way of assessing the quality of the data. But this approach has shortcomings. Entropy is a complex measure of uncertainty and therefore unintuitive. In addition we do not know how outliers might affect the ranking of the stations. Since

the entropy approach is not based on a predictive model it does not yield a method which could actually be used to predict the ion concentrations of the stations which would be dropped from the network.

To address these shortcomings, a different approach is taken here to determine the relative quality of the environmental data. In this approach, a station is considered to yield high quality data if its observations are difficult to predict from observations of other stations. We do not mean to suggest that we are discrediting the entropy approach but rather we are aiming at developing a better understanding of it.

CHAPTER 2

METHODS FOR PREDICTING ONE STATION'S RECORDS FROM THE OTHERS

2.1 Introduction

A multiplicity of plausible objectives can be foreseen for any data collection network, and at the same time some important future uses of the data may not be foreseen. This poses a dilemma as quality represents fitness for intended use (see Caselton, Kan and Zidek, 1990). As noted in the Introduction, to circumvent this difficulty, Wu and Zidek (1992) use an entropy based approach to assess the relative quality of the data produced by each station in a data gathering network, specifically that network which is the the subject of our study.

The goal of the present study is similar to that of Wu and Zidek (1992) except that we use a concept other than entropy to define data quality. Like Wu and Zidek (1992), we look at the data quality for each station as the amount of additional information provided by the data in that station. However we define the amount of information as the extent to which the data from a particular station can be predicted from that of the other stations. A station whose data are hard to predict is considered as adding much information to the network. We would interpret such a station's data as being of high quality. By a similar argument, stations whose data are easily predicted are thought to add little information, which we interpret to mean the data are of low quality. We could argue that, if there is a need to reduce the size of the network, the stations with low quality data should be dropped out of the network first.

The methods used in this study to predict one station's rainfall chemistry records from the others are ordinary regression, regression using the Bayesian approach and Stone's cross-validatory assessment method (See Stone, 1973). Cross validation is used

to assess the performance of all the methods.

In all three methods, regression models are constructed to predict the ion concentration of station j from the ion concentrations of the remaining stations in a cluster. Months are considered as replicates. For instance, the ordinary regression method results in a regression model for each month $i = 1, \ldots, 48$ and each station $j = 1, \ldots, p$, where p is the size of the cluster. For fixed i and j the regression model is constructed using the remaining 47 months as replicates. Thus, there are p parameters to be estimated from 47 "replicates". The values of p range from 2 to 47 and, in some of the larger clusters, p is close to and sometimes even greater than the number of months with complete observations. For example, in one of the sulfate clusters with $p = 36$, there are only 19 months for which all the station values are available. It is not possible to perform an ordinary regression analysis when there are only $19 - 1 = 18$ cases available to estimate 36 parameters. So our analysis is done with missing observations replaced by their estimates at the outset: for each cluster, we produce an X matrix of logarithmically transformed ion concentrations with no missing entries. These completed data sets are used for all methods to achieve consistency.

Our strategy for estimating missing observations is discussed in the next section. The last three sections of this chapter discuss in detail the three methods used to predict a station's rainfall chemistry from the others.

2.2 Estimation of Missing Observations.

Various methods for estimating missing observations have been proposed by different authors. Afifi and Elashof (1961) suggest filling in the missing observations for each variable by that variable's mean. Another method uses regression instead. A variable with missing observations is regressed on the other variables in the study, using only complete cases. Regression is done separately for each variable with missing values.

The estimated regression model is used to impute the missing values. This method is discussed by Buck (1960) as well as Afifi and Elashof (1961). Stein and Shen (1991) regressed the logarithm of sulfate concentrations on the amount of precipitation and the month of the year when precipitation occurred. The model fitted by least squares was used to impute the missing sulfate values.

A modified regression strategy was used in this study to impute the missing values. Each station with missing values was regressed on the other stations using only complete records. But as we saw above, in some of the clusters the number of complete records was less than the number of parameters to be estimated. So the regression method needed some modifications before it could be applied to our data. These modifications are found in a BMDP program, which estimates missing values. The program is called "Description and estimation of missing data" and is abbreviated as "PAM".

This method uses stepwise regression to select the variables to be used. First, the variable most correlated with the variable with missing values is chosen to enter into the regression equation. If the chosen variable meets the "F-to-enter" criterion (explained below), then the next variable is chosen. The variable chosen next is the one with the highest partial correlation with the variable with missing values, with the partial correlation conditional on the variable already used in the equation. This variable must also meet the "F-to-enter" criterion. Additional variables are chosen in the same manner until all variables which meet the "F-to-enter" criterion have been used. If, during stepwise regression, no variables satisfy the criterion for admission into the regression, the mean of the variable with missing values is used to fill in that variable's missing values.

The "F-to-enter" criterion is motivated by an approximate test of the coefficient of any predictor variable. That is, the square of the ratio of the predictor's regression

coefficient to its standard error is approximately distributed as an F statistic with one degree of freedom for the numerator. The square of this ratio is compared with the pre-set "F-to-enter" limit. If the square is greater than this limit, the variable has satisfied the "F-to-enter"criterion. For further details on this program, refer to BMDP Statistical Software, 1983 edition, page 217 and Frane (1978b).

2.3 Ordinary Regression Using Cross-Validatory Assessment

Ordinary regression using cross-validatory assessment is one of the methods used in this study to assess how well each station's rainfall chemistry can be predicted from the chemistry records of other stations. In this study the sample is divided into two parts. The first part is used for estimation while the second part is for assessment. The size of the estimation subsample is taken to be $n - 1$ (with $n = 48$) and that of the prediction subsample to be 1. These are the same subsample sizes used by Stone (1973).

We proceed as follows. One station is selected and its data values designated as the predictands. The remaining station records provide the predictors. Once a station is fixed, we set aside one month's record from among the $n = 48$ monthly records, and treat this as a "future" month for prediction. Now the data for the remaining 47 months are used as a "training set"; a linear model is fitted by ordinary least squares, using the 47 months as replicates. The estimated regression equation is then used to predict the ion concentration of the future month for the designated station, with the remaining stations' concentrations for that month as predictors. The selected month is now replaced by another and the process repeated until all 48 months have played the role of the future month. In this way we obtain 48 predictions and 48 prediction errors for the designated station. That station is now replaced by another from the cluster and the whole exercise repeated. A new station now provides the predictands and this continues until all stations in the cluster have played the designated role. At this point

we can assess the efficacy of ordinary regression and the relative difficulty of predicting the records of the various stations from the others.

To state our procedures more precisely let p be the number of stations in a cluster and $n = 48$ the number of months in the study. Further let :

x_{ij} be the logarithmically transformed ion concentration for the i^{th} month in the j^{th} station,

$X = (x_{ij})$, $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, p$, be the $n \times p$ observation matrix,

$X_{i.} = (x_{i1}, \ldots, x_{ip})$, $i = 1, \ldots, 48$, and $X_{.j} = (x_{1j}, \ldots, x_{nj})^t$, $j = 1, \ldots, p$,

$X^{-.j}$ be the matrix X with column vector $X_{.j}$ deleted,

$X^{-i.}$ be the matrix X with row vector $X_{i.}$ deleted,

X^{-ij} be the matrix X with both column and row vectors $X_{.j}$ and $X_{i.}$ deleted,

$X_{.j}^{-i}$ be the column vector $X_{.j}$ with the i^{th} month deleted,

$X_{i.}^{-j}$ be the row vector $X_{i.}$ with the j^{th} station deleted,

$X^{-ij,1}$ be the matrix X^{-ij} with a column of ones added as the first column,

$X_{i.}^{-j,1}$ be the row vector $X_{i.}^{-j}$ with one added as a first element.

For fixed j and $i = 1, 2, \ldots, 48$ the dependence of $X_{.j}^{-i}$ on $X^{-ij,1}$ is modeled as

$$X_{.j}^{-i} = X^{-ij,1}\underset{\sim}{\beta} + \underset{\sim}{U} \tag{3}$$

where $X_{.j}^{-i}$ is a 47×1 column vector, $X^{-ij,1}$ is a $47 \times p$ matrix, $\underset{\sim}{\beta}$ is the $p \times 1$ vector of regression coefficients and $\underset{\sim}{U}$ is a 47×1 column vector of random errors. Denote the estimated regression coefficients (which depend on i and j) and predicted value of x_{ij} by $\hat{\underset{\sim}{\beta}}$ and $XEST_{ij}$, respectively. Then $\hat{\underset{\sim}{\beta}}$ is multiplied by $X_{i.}^{-j,1}$ to get $XEST_{ij}$. The prediction error made in predicting x_{ij} is given by $x_{ij} - XEST_{ij}$ and denoted by PE_{ij}.

Both $XEST_{ij}$ and PE_{ij} are on a logarithmically transformed scale. Each station yields a vector of 48 elements for both predicted values and prediction errors, one element for each month. The average of the squared prediction errors for each station is used to assess how well each station's rainfall chemistry can be predicted from that of the other stations. For fixed station, j, let $\mathcal{I}_j$ be the set of indices for which x_{ij} was not imputed, that is, the set of i's for which x_{ij} was not missing in the original data set. Then, the average of the squared prediction errors for station j is calculated using prediction errors only for months i in $\mathcal{I}_j$. Since the x_{ij} values for the months with $i \notin \mathcal{I}_j$ were imputed using regression, the prediction errors from these months are omitted from the average to avoid misleadingly small estimates of the prediction error. Now if we denote the average squared prediction error of station j by APE_j, then

$$APE_j = (\sum_{i \in \mathcal{I}_j} PE_{ij})/(48 - k_j), \tag{4}$$

where k_j is the number of months for which station j had missing data. Stations are ranked using APE. Within a cluster the station with the minimum value of APE is ranked first and has the lowest quality data according to the criterion described above. The station with the highest quality data is the one with maximum APE and is ranked last.

2.4 Regression Using a Bayesian Approach

In the previous section, a least squares approach is used to estimate the unknown regression coefficients in predicting one station's ion concentrations from the others. But if prior information about the regression coefficients is available, then this information should be exploited to find improved regression coefficient estimates. In this section, we assume that such prior information exists. Specifically, when station j and month i are fixed, the model

$$X_{corr.j}^{-i} = X_{corr}^{-ij}\underset{\sim}{\beta} + \underset{\sim}{U}, \tag{5}$$

is assumed, where X_{corr} denotes the matrix of data values centred about station averages. Prior knowledge on the regression coefficients, $(\beta_1, \ldots, \beta_{p'})$ is expressed in terms of a density which is exchangeable, where $p' = p - 1$. That is, prior knowledge of the regression coefficients would be unaltered by any permutation of the suffixes. This implies that our opinion of β_j is the same as that of β_l.

This idea of exchangeability derives from the work of de Finetti (1964). Lindley and Smith (1971) apply this idea and give explicit expressions for the Bayesian estimates. They use the mixture approach to construct an exchangeable prior distribution of the parameters of interest. This mixture approach is supported by Hewitt and Savage (1955). To give a simple example from Lindley and Smith (1971), suppose $E(y_i|\underset{\sim}{\theta}) = \theta_i$ and $E(\theta_i) = \mu$, with $var(y_i|\underset{\sim}{\theta}) = \sigma^2$ and $var(\theta_i) = \tau^2$. If it is assumed that the θ_i's are exchangeable, then, P, the prior distribution on $\underset{\sim}{\theta}$, is of the form

$$P(\underset{\sim}{\theta}) = \int \prod_{i=1}^{n} P(\theta_i|\mu) d\phi(\mu), \tag{6}$$

where $P(\cdot|\mu)$ and $\phi(\cdot)$ are arbitrary measures. In the language of Lindley and Smith (1971), this example is a two stage-model. If in turn, ϕ depends on unspecified "hyper-parameters", then we can have a three stage-model. The choice of the number of stages to be used depends on the individual.

Lindley and Smith (1971) consider the linear regression model of the form,

$$\underset{\sim}{Y} = T\underset{\sim}{\beta} + \underset{\sim}{U}, \tag{7}$$

where the prior opinion on $\underset{\sim}{\beta}$ is exchangeable. We modify their approach by first "centering" the data about their mean to eliminate the need for an intercept coefficient in Equation 7. Because of the exchangeability assumption, we can not include the intercept in the regression, since its inclusion would make the exchangeability assumption unrealistic. To be precise, after setting aside the values for the "future month", we subtract

each station's mean (based on 47 months) from each station's concentrations. Then we work with these corrected values. This forces the regression line to pass through the origin. We apply Lindley and Smith's method (1971) with the regression model of Equation 5, where $X^{-i}_{corr.j} = X^{-i}_{.j} - \underset{\sim}{1}(\sum_{k \neq i} x_{kj})/47$ and $\underset{\sim}{1}$ is a 47×1 vector all of whose elements are 1. Each station's values in X^{-ij}_{corr} are obtained from the corresponding station's values in X^{-ij} in the same way as $X^{-i}_{corr.j}$. The resulting Bayesian estimate of $\underset{\sim}{\beta}$, $\underset{\sim}{\beta}^*$, is then used to calculate $\hat{x}_{ij}$, the prediction of station j during month i, by,

$$\hat{x}_{ij} = (\sum_{k \neq i} x_{kj})/47 + X^{-j}_{corri.}\underset{\sim}{\beta}^*, \tag{8}$$

where $X^{-j}_{corri.}$ is obtained from $X^{-j}_{i.}$ by subtracting from each element of $X^{-j}_{i.}$ the appropriate station average.

Suppose

$$X^{-i}_{corr.j}|\underset{\sim}{\beta}, \xi, \sigma^2, \sigma^2_\beta \sim N(X^{-ij}_{corr}\underset{\sim}{\beta}, \sigma^2 I_{n'}), \tag{9}$$

where $I_{n'}$ is the $n' \times n'$ identity matrix and $n' = 47$. Assume $\underset{\sim}{\beta}^t = (\beta_1, \ldots, \beta_{p'})$ is exchangeable and that

$$\beta_j|\xi, \sigma^2_\beta \sim N(\xi, \sigma^2_\beta). \tag{10}$$

Assuming vague prior knowledge for ξ, Lindley and Smith (1971) give an explicit expression for the Bayesian estimate as

$$\underset{\sim}{\beta}^* = \{I_{p'} + k((X^{-ij}_{corr})^t X^{-ij}_{corr})^{-1}(I_{p'} - (J_{p'})/p')\}\underset{\sim}{\hat{\beta}}, \tag{11}$$

where $k = \sigma^2/\sigma^2_\beta$, $\underset{\sim}{\hat{\beta}}$ is the usual least squares estimate, p' is the number of parameters, and $J_{p'}$ is a matrix of dimension $p' \times p'$, all of whose elements are 1.

In practice σ^2 and σ^2_β will be unknown, and they can be estimated from the data. To estimate σ^2 and σ^2_β, it is now assumed that σ^2 and σ^2_β, which are independent of $\underset{\sim}{\beta}$ and ξ, are independently distributed as

$$\nu\lambda/\sigma^2 \sim \chi^2_\nu, \tag{12}$$

$$\nu_\beta \lambda_\beta / \sigma_\beta^2 \sim \chi^2_{\nu_\beta}, \tag{13}$$

(see Lindley and Smith, 1971), where ν, λ, ν_β and λ_β are prior parameters. The joint distribution of $X^{-i}_{corr.j}$, $\underset{\sim}{\beta}$, σ^2 and σ^2_β is given by

$$\begin{aligned}(\sigma^2)^{-1/2n} exp\{-(X^{-i}_{corr.j} - X^{-ij}_{corr}\underset{\sim}{\beta})^t(X^{-i}_{corr.j} - X^{-ij}_{corr}\underset{\sim}{\beta})/(2\sigma^2)\} \times \\ (\sigma^2_\beta)^{-1/2p} exp\{-(\underset{\sim}{\beta} - \underset{\sim}{1}\xi)^t(\underset{\sim}{\beta} - \underset{\sim}{1}\xi)/(2\sigma^2_\beta)\} \times \\ (\sigma^2)^{-\frac{1}{2}(\nu+2)} exp\{-\nu\lambda/(2\sigma^2)\} \times (\sigma^2_\beta)^{-\frac{1}{2}(\nu_\beta+2)} exp\{-\nu_\beta\lambda_\beta/(2\sigma^2_\beta)\}.\end{aligned} \tag{14}$$

Integrating the joint distribution with respect to ξ, one calculates the posterior distribution of $\underset{\sim}{\beta}$, σ^2 and σ^2_β, which is proportional to

$$\begin{aligned}(\sigma^2)^{-1/2(n+\nu+2)} exp[-\{\nu\lambda + (X^{-i}_{corr.j} - X^{-ij}_{corr}\underset{\sim}{\beta})^t(X^{-i}_{corr.j} - X^{-ij}_{corr}\underset{\sim}{\beta})\}/(2\sigma^2)] \\ \times(\sigma^2_\beta)^{-1/2(p+\nu_\beta+1)} exp[-\{\nu_\beta\lambda_\beta + \sum_{k=1}^{p'}(\beta_k - \beta_.)^2\}/(2\sigma^2_\beta)],\end{aligned} \tag{15}$$

where $\beta_. = (\sum_{k=1}^{p'} \beta_k)/(p')$. Using the results in Equation 11 and the modal equations for σ^2 and σ^2_β from the posterior distribution, we get the Bayes estimates of $\underset{\sim}{\beta}$, σ^2 and σ^2_β :

$$\begin{aligned}\underset{\sim}{\beta}^* &= \{I_{p'} + k^*((X^{-ij}_{corr})^t X^{-ij}_{corr})^{-1}(I_{p'} - (J_{p'})/p')\}\underset{\sim}{\hat{\beta}}, \\ s^2 &= \{\nu\lambda + (X^{-i}_{corr.j} - X^{-ij}_{corr}\underset{\sim}{\beta}^*)^t(X^{-i}_{corr.j} - X^{-ij}_{corr}\underset{\sim}{\beta}^*)\}/(n' + \nu + 2), \\ s^2_\beta &= \{\nu_\beta\lambda_\beta + \sum_{k=1}^{p'}(\beta^*_k - \beta^*_.)^2\}/(p' + \nu_\beta + 1),\end{aligned} \tag{16}$$

where $k^* = s^2/s^2_\beta$. Lindley and Smith (1971, p17) suggest that the parameters ν and ν_β may well be small positive numbers and that, in any case, the solution is very insensitive to changes in these numbers, so that they may be set to zero. In the example in Lindley and Smith (1971), the values of ν and ν_β were set to zero as well as was the starting value of k^*. The resulting equations were solved iteratively, starting with initial values for s^2 and s^2_β or by setting the starting value of k^* to zero. The initial value of $\underset{\sim}{\beta}^*$ was

found via (16) and then used to find the next values of s^2 and s_β^2, and so on. However, if ν and ν_β are set to zero and if during iteration, the values of the estimates of the β_l's become close to each other, then s_β^2 becomes very close to zero. Hence $k^* = s^2/s_\beta^2$ becomes very big. This can result in an overflow in the computer calculations. From (16), it can be seen that $\underset{\sim}{\beta}^* \to 0$ as $k^* \to \infty$. Thus if a small estimate of $\underset{\sim}{\beta}$ is used in our prediction model, Equation (8), the prediction will be close to the the station's average. Because of the problem of overflow, we suggest two alternatives of estimating $\underset{\sim}{\beta}$. The first alternative is to set $\nu_\beta \lambda_\beta$ in the last Equation of (16) to a small number such that ν_β is negligible and can be ignored in the denominator of the same expression. This modification keeps the estimate, s_β, bounded away from zero in the iterations. The second alternative is to not iterate, but rather to pick a value of k^* and use it in the first Equation of (16) to calculate the estimate, $\underset{\sim}{\beta}^*$.

Because we do not have grounds to prefer one of the alternatives over the other, both are used. In the first alternative, which we will call Bayesian regression alternative (1), the initial value of k^* is chosen to be zero, as in Lindley and Smith (1971). In subsequent iterations, $\nu_\beta \lambda_\beta$ is set equal to 0.1, with ν_β negligible, so that

$$s_\beta^2 = \{0.1 + \sum_{k=1}^{p'} (\beta_k^* - \beta_\cdot^*)\}/(p'+1).$$

In the second, non-iterative alternative which we will call Bayesian regression alternative (2), we take

$$s^2 = (X_{corr.j}^{-i} - X_{corr}^{-ij}\underset{\sim}{\hat{\beta}})^t (X_{corr.j}^{-i} - X_{corr}^{-ij}\underset{\sim}{\hat{\beta}})/(n'+2)$$

$$s_\beta^2 = \sum_{k=1}^{p'} (\hat{\beta}_k - \hat{\beta}_\cdot)^2/(p'-1).$$

However for a cluster of size two the second alternative can not work, because after fixing a station, there is only one station as a predictor, and thus $p' = 1$, and s_β^2 is undefined.

These two approaches are applied to our data to give the Bayesian estimates of the regression coefficients for each data set constructed by fixing station j and month i, and using the remaining 47 months to construct a model for the prediction of the ion concentration of station j. For fixed station j and month i, we get the Bayesian estimate $\underset{\sim}{\beta}^*$. This $\underset{\sim}{\beta}^*$ is used to calculate the predicted value of x_{ij}, denoted by $\hat{x}_{ij}$. We get the prediction error made by predicting x_{ij}, namely $x_{ij} - \hat{x}_{ij}$. For each station j, we obtain the vector of 48 predicted values and prediction errors, one for each month. The average of the squared prediction errors for station j, calculated using only the months which were not imputed, is used to give the assessment of the station's predictability. For each cluster, stations are ranked using the average of the squared prediction error. The station with the minimum average squared prediction error is ranked first, and is considered as having the lowest data quality. Similarly, the station with the maximum average squared prediction error is ranked last and considered as having the highest data quality. The results are included in the Appendix and the discussion is in Chapter 3.

2.5 Stone's Procedure

Both the ordinary regression method and regression based on a Bayesian procedure, for predicting rainfall chemistry of a particular station from the chemistries of the other stations, were discussed in the previous two sections. These methods were assessed by cross-validation. If it is accepted in advance that cross-validation will be used in this way, then the ordinary least squares and the Bayesian predictors can be modified to do well in the intended assessment. The resulting predictor differs markedly from that produced by the previous two methods. It is due to Stone (1973) and we include it in our study as a competitor to the first two methods.

Stone's procedure begins by choosing a statistical predictor, which is a function of the data, $(y_i, \underset{\sim}{t}_i)$ $i = 1, \ldots, n$, and a parameter denoted by α. Here $\underset{\sim}{t}_i^t = (t_{i1}, \ldots, t_{i\kappa})$.

The parameter, α, is estimated from the data by cross-validation, using a loss function. This loss function is also used to calculate C^+, an assessment of the prediction efficiency. The score is constructed by calculating $\hat{y}_i$, a statistical predictor of y_i based on the data set (y_l, t_l), $l = 1, \ldots, n$, $l \neq i$. The assessment score, C^+, is the average of the losses incurred from estimating y_i by $\hat{y}_i$.

We apply Stone's method to each station in a cluster. That is, for each fixed station j, we take $n = 48$, $\kappa = p - 1$, $y_i = x_{ij}$ and $\underset{\sim}{t}_i^t = X_{i\cdot}^{-j}$, the vector of ion concentrations of the remaining stations in month i. Stone's procedure yields a linear model for predicting station j's ion concentrations from the other stations', and a score, C_j^+, which assesses the predictability of station j. These scores are then used to rank the stations within a cluster.

In Section 3 of his paper, Stone (1973) gives a number of statistical predictors which can be used in different problems. In the case of the present study, the statistical predictor of Stone's Example 3.3 is appropriate. Letting $s = \{1, \ldots, n\}$ the statistical predictor based on $(y_i, \underset{\sim}{t}_i)$, $i \in s$, is given by:

$$\hat{y}(\underset{\sim}{t}_i, \alpha, s) = \alpha\bar{y} + (1 - \alpha)(\bar{y} + \sum b_k(t_{ik} - \bar{t}_{\cdot k})), \tag{17}$$

where $\bar{y} = \frac{1}{n}\sum_1^n y_i$, and the $\bar{t}_{\cdot k} = \frac{1}{n}\sum_{i=1}^n t_{ik}$, and the b_k's are the estimated regression coefficients in the least squares multiple regression of y on $\underset{\sim}{t}$. The parameter, α, in the statistical predictor, is estimated via cross-validation, using squared error loss. That is, α is chosen to minimize :

$$C(\alpha) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}(\underset{\sim}{t}_i, \alpha, s^{-i}))^2 \tag{18}$$

where $s^{-i} = \{1, \ldots, n\}/\{i\}$. The value of α so obtained is denoted by $\alpha^+(s)$, and the resulting model is Equation 17 with $\alpha = \alpha^+(s)$.

Stone gives the explicit form of $\alpha^+(s)$ corresponding to the selected statistical predictor and loss function as

$$\alpha^+(s) = \{\sum_{i=1}^{n}[\frac{r_i^2}{(1-A_{ii})^2} - \frac{nr_i(y_i-\bar{y})}{(n-1)(1-A_{ii})}]\}/\{\sum_{i=1}^{n}[\frac{r_i}{(1-A_{ii})} - \frac{n(y_i-\bar{y})}{(n-1)}]\} \qquad (19)$$

where r_i is the i^{th} residual in the least squares multiple regression using the data $(y_i, \underset{\sim}{t}_i)$, $i \in s$, and

$$A = T(T^tT)^{-1}T^t,$$

where T is the design matrix corresponding to this regression.

The cross-validatory assessment employs C^+, which is calculated as follows. For each $i = 1, \ldots, n$, the statistical predictor of Equation 17 is constructed as described above, but using the reduced data set $(y_k, \underset{\sim}{t}_k)$, $k \in s^{-i}$. Thus for each i, we have a cross-validatory choice of α, $\alpha^+(s^{-i})$, and an estimate of y_i, $\hat{y}(\underset{\sim}{t}_i, \alpha^+(s^{-i}), s^{-i})$. The assessment of predictability is given by

$$C^+ = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}(\underset{\sim}{t}_i, \alpha^+(s^{-i}), s^{-i}))^2. \qquad (20)$$

This statistical predictor and assessment procedure are applied in the present study to each station. The station whose predictability score, C_j^+, is lowest is the most easily predicted, so it is considered to have low quality data according to this assessment procedure. Similarly a station with a larger value of C_j^+ is hard to predict, and thus is considered to be producing data of high quality. Stations are ranked from low to high according to this definition of their data quality. Discussion of the results is given in Chapter 3. The rankings are given in the Appendix.

CHAPTER 3

ASSESSMENT OF THE NETWORK AND THE AGREEMENT OF THE METHODS

The methods described in Chapter 2 were used to assess the quality of the data described in Section 2 of Chapter 1, obtained over the period, 1983-86; and the results are presented in this section. The data used in this study were collected by the National Deposition/National Trend Network (NADP/NTN) which we discussed in Section 2 of Chapter 1. The NADP/NTN network stations used in this study are tabulated in the Appendix. For reasons given in Section 1 of Chapter 2, we use the data with missing values replaced by their estimates as described in Section 2 of Chapter 2. After filling in the missing values by their estimates, we have a total of 48 volume weighted monthly average concentrations for each station and ion, and these average concentrations are used in all the methods.

In Section 2 of Chapter 1, we discussed the nature and the effect of acid deposition. Our study involves the concentrations of three ions, namely: hydrogen, sulfate and nitrate. The data were logarithmically transformed, to achieve a more nearly Gaussian data distribution and to be consistent with the earlier work done on the same data set. The clusters from the study by Wu and Zidek (1992) are used.

Our results for sulfate ion concentrations will be discussed in detail to illustrate our findings. This ion is selected for detailed consideration, so that one can compare our results with those from the entropy based analysis of Wu and Zidek (1992) where the same ion was selected for detailed discussion. Alongside this focused discussion we shall be commenting generally on all the ions and their clusters.

Data for sulfate ion concentrations yield 3 station clusters with 37, 36 and 8 stations. Tables A1.2(a)-A1.2(c) in the Appendix give the ranks of the stations for each cluster

determined by the methods used in our study and those given by entropy analysis. Also included in the tables are the average squared prediction errors, an index of quality used to rank stations. The corresponding measure used to rank stations in entropy analysis is not included in the tables, because it is not comparable to average prediction error.

To focus our discussion, consider the third sulfate cluster which contains 8 stations, with identification codes 037a, 059a, 061a, 074a, 078a, 271a, 281a, and 354a. The corresponding station names can be found in the Appendix. Table 3.1 below (identical to Table A1.2(a) in the Appendix, but reproduced here for convenience) contains the ranks and average prediction errors for the 8 stations in the cluster. Since this cluster contains 8 stations, a rank of 8 corresponds to the best station, a rank of 7 to the second best station, and so on. Four of the stations in the cluster (half the cluster size) are given the same ranks by all the methods. The station with identification code 037a is ranked first by the three methods used here, as well as by the entropy analysis. For this station, the average squared prediction errors produced by ordinary regression, regression using the Bayesian approach and Stone's cross-validatory assessment method are respectively, 0.1484, 0.2242 and 0.1483. This means that this station (Glacier National Park, Montana) has the lowest quality (most easily predicted) data according to the four methods. If there were a need to reduce the size of the network, this station might be considered first for possible termination. Two other stations in the cluster compete for "best". Identification codes for these stations are 281a and 354a, with respective names Bull Run, Oregon, and St. Mary Ranger Station, Montana. Bull Run, Oregon, is ranked best by the ordinary regression method and Stone's procedure, while both the Bayesian alternative approaches and entropy rank it second best. St. Mary Ranger Station, Montana, is ranked best by both the Bayesian alternatives approaches and entropy, while ordinary regression and Stone's procedure rank it second best.

Table 3.1: Average Squared Prediction Errors and Ranks in One of Sulfate Cluster

station	methods								
	Regression		Bayesian 1		Bayesian 2		Stone		Entropy
code	APE	rank	APE	rank	APE	rank	APE	rank	rank
037a	0.1484	1	0.1343	1	0.1334	1	0.1483	1	1
059a	0.3091	5	0.2683	5	0.2711	5	0.2836	5	4
061a	0.2537	4	0.2354	4	0.2403	4	0.2279	4	5
074a	0.4518	6	0.3693	6	0.3812	6	0.4096	6	6
078a	0.1796	2	0.1635	2	0.1659	2	0.1610	2	2
271a	0.2498	3	0.1979	3	0.1983	3	0.2276	3	3
281a	0.5258	8	0.4044	7	0.4604	7	0.5398	8	7
354a	0.5053	7	0.4773	8	0.4767	8	0.4472	7	8

In general, the four different methods do not give the same best station. However, as can be seen from the table above, the difference is not very important, since a station ranked best by one method is either ranked best or second best or third best by the other methods. Similarly, the ranks for the intermediate stations do not have significant differences. There is no cluster where a station is identified as the best by one method, and the poorest by the other methods. If it were literally necessary to select the best single station it would be necessary to pay careful attention to its selection. One might well face a decision problem, and other nonstatistical issues might be invoked to resolve the conflict. For example, the geographical positions of the stations might be taken into consideration.

It is not unusual for different methods, adopted for a single purpose, to give different results, or for different judges to give different ranks to various contestants. But it is important that there be a reasonably strong association between the ranks given by the

different methods. Below we use association measures to give a more precise assessment of the agreement of the ranks from the five methods.

We use association measures to assess the degree of agreement between pairs of ranking methods and among all five ranking methods. This is done within each cluster. The ten pairwise associations we consider are for ordinary regression with the Bayesian regression alternative (1) approach, ordinary regression with the Bayesian regression alternative (2) approach, ordinary regression with the Stone's procedure, ordinary regression with the entropy approach, the Bayesian regression alternative (1) approach with the Bayesian regression alternative (2) approach, the Bayesian regression alternative (1) approach with the Stone's procedure, the Bayesian regression alternative (1) approach with the entropy approach, the Bayesian regression alternative (2) with Stone's procedure, Bayesian regression alternative (2) with entropy and Stone's procedure with entropy approach. We test the null hypothesis that the rankings are random against the alternative hypothesis that a direct association exists between the ranking methods.

For pairwise associations, Spearman's coefficient of rank correlation is used. The test statistic is $R = 1 - 6\sum_{i=1}^{p} D_i^2/(p^3 - p)$, where D_i is the the difference between the two ranks given the i^{th} station, and p is the number of stations in the cluster. The value of R lies between -1 and 1, where a value of -1 means perfect disagreement or inverse association and a value of 1 means perfect agreement or direct association, which is of interest in our study. A value of zero indicates there is no association, that is neither agreement nor disagreement. For $p \leq 30$ exact p-values for the calculated values of R are given in Table I of Gibbons (1976) while for larger values of p , $p > 30$, an approximate normal distribution is used to calculate the p-values. ($R\sqrt{p-1}$ is approximately standard normal under the null hypothesis)

For the five different ranking methods, Kendall's coefficient of concordance is used to

test the same hypothesis of no association. Three different test statistics are available. The first of these, denoted by S, measures the departure from lack of agreement and is given by,

$$S = \sum_{j=1}^{p} [C_j - k(p+1)/2]^2, \tag{21}$$

where C_j is the sum of rankings of the j^{th} station and $k = 5$ is the number of ranking methods. This quantity is expected to be small if there is no agreement and big if a positive association exists. But it is difficult to have an intuition about the size of S, so a second test statistic denoted by W is used. W is a relative measure of association, defined by $W = S/S^*$, a ratio of the observed measure of departure from lack of agreement, S, to S^*, where S^* is given by

$$S^* = \sum_{j=1}^{p} [jk - k(p+1)/2]^2. \tag{22}$$

S^* is the value of S under perfect agreement. The value of W lies between 0 and 1, where a value of 0 means no association and a value of 1 means perfect agreement. Accordingly, large values of W call for rejection of the null hypothesis in favor of the alternative. Since the test statistic, S, is a monotonically increasing function of W, and is large when W is large and zero when W is zero, the appropriate p-values are the right tail probabilities. Table K of Gibbons (1976) gives exact right-tail probabilities for S for $p = 3, k \leq 8$, and $p = 4, k \leq 5$.

For combinations of p and k that are not covered by that table, an equivalent test statistic to S and W, denoted by Q, is used. Either of the following two expressions can be used to calculate Q:

$$\begin{aligned} Q &= k(p-1)W, \\ Q &= 12S/kp(p+1). \end{aligned} \tag{23}$$

The distribution of this test statistic, Q, can be approximated by the chi-square distri-

bution with $p-1$ degrees of freedom for large k. As with the test statistic S, Q is also a monotonically increasing function of W, so the appropriate p-values are also the right tail probabilities. Details of the association test procedures for both pairwise and any number, k, of ranking methods is found in Gibbons (1976).

The association measures for all the clusters and all the ions are given in Tables A2.1-A2.3 in the Appendix. The entries in the table are R for pairwise associations or W for all five ranking methods, Q or S for all four ranking methods, depending on which statistic is used to calculate the p-value and, the p-values. Table 3.2 below (identical to Table A2.2(c) in the Appendix but given a different label to be consistent within this chapter) contains the association measures for the third sulfate cluster with 8 stations.

Table 3.2: Association measures for the third sulfate cluster

methods being compared	R or W	Q	p-value
regression versus Bayesian 1	0.976	-	0.000
regression versus Bayesian 2	0.976	-	0.000
regression versus Stone	1.000	-	0.000
regression versus entropy	0.952	-	0.001
Bayesian 1 versus Bayesian 2	1.000	-	0.000
Bayesian 1 versus Stone	0.976	-	0.0000
Bayesian 1 versus entropy	0.976	-	0.000
Bayesian 2 versus Stone	0.976	-	0.000
Bayesian 2 versus entropy	0.976	-	0.000
Stone versus entropy	0.952	-	0.001
all methods	0.981	34.3333	0.000

The values of R and W for this cluster are greater than 0.9. Also the p-values are less than or equal to 0.001. These results give evidence for the rejection of the null

hypothesis of random ranking in favor of the alternative hypothesis. That is, in this cluster the ranks of the stations given by the all five methods have a strong agreement. Looking at the association measure, W, for all rankings, in all the clusters, we see that, for most of the clusters, W is greater than 0.8 and the p-value is less than 0.001 . This indicates good agreement of the ranks from all the methods. Results for nitrate show a different pattern from that discussed above. The association measures for nitrate clusters are relatively low. For example, for one of nitrate clusters with 41 stations W is 0.5. But the p-value for this cluster is small (0.00001) so this gives grounds to reject the random ranking hypothesis in favor of the alternative. Generally small clusters of size less than five have large association measures but relatively large p-values. But this cannot be used to support the null hypothesis of random ranking, since these clusters have inadequate sample sizes to yield a reliable conclusion.

This study was motivated by the results of the entropy analysis. The entropy approach seems to work well in ranking stations on the basis of data quality, but it is not intuitive. So we seek an intuitive method which gives ranks close to those from entropy analysis; it could be used in combination with the entropy approach to get a better understanding of the data. The scatterplots of the p-values against cluster sizes and relative measures, R, against cluster sizes are used to find out which of the methods used here is in close agreement with the entropy analysis and to see if agreement depends on cluster size. The plots are done ion-by-ion. Scatterplots for all the ions are included in the Appendix. Figures 3.1(a) and Figure 3.1(b) below are scatterplots for respectively, the relative measure of agreement, R, against cluster sizes and the p-values against cluster sizes for sulfate ion.

From Figure 3.1(a) we see that the line corresponding to the Bayesian alternative (1) and entropy is above the other lines and the next highest line is the one correspond-

ing to Stone's procedure and entropy. This means that the agreement, as measured by R, between the Bayesian alternative (1) and entropy is higher than for the others. In addition, the corresponding line in Figure 3.1(b) is below the other lines. This indicates that the p-values for this comparison are smaller than for the others. These observations suggest that the ranks produced by the Bayesian alternative (1) approach agree with those produced by entropy more often than the others. The same phenomenon is observed with the hydrogen ion rankings except for one cluster in which the ordinary regression and entropy comparison has a bigger value of R and a lower p-value. But this one cluster among the six clusters of hydrogen cannot change our conclusion. That is, even for hydrogen, the Bayesian alternative (1) approach and Stone's procedure agree strongly with entropy. Nitrate gives exceptional results. For this ion Stone's procedure seems to rank last in agreement with entropy, but the Bayesian alternative (1) once again shows strong agreement with entropy.

The scatterplots indicate that there is no consistent relationship between agreement and cluster size.

The need for diagnostic checking is one of the reasons which motivated this study of various approaches to assessing data quality. We need to know more about the predicted values and the prediction errors. We acquire this knowledge by comparing the boxplots of the predicted values to those of the observed values, and by studying the boxplots of the prediction errors.

For each cluster, we look at two sets of boxplots. One set of boxplots shows the observed and predicted values from all four methods used in this study. For easy reading, we frame the five boxes for each station separately with the station identification code above the frame. The first boxplot in each frame is for the observed values, while the second, third, fourth and fifth represent the predicted values from respectively, ordinary

regression, regression using the Bayesian alternative (1) approach, regression using the Bayesian alternative (2) approach and Stone's procedure. The second set of boxplots shows the prediction errors from the four methods used. The same format as above is repeated. The four consecutive boxes in each small frame represent the prediction errors for each given station, where the first boxplot in a group is for the prediction errors from the ordinary regression method, the second and third boxes are, respectively, for alternative 1 and alternative 2 of the Bayesian approach while the last boxplot represents the prediction errors from Stone's procedure.

Boxplots for all the clusters are included in the Appendix. Both sets for the third sulfate cluster are included in this chapter for easy reference, and are labelled 3.2(a) and 3.2(b) for consistency within the chapter. Figure 3.2(a) is for the boxplots of the observed and predicted values, while figure 3.2(b) is for the boxplots of the prediction errors.

From Figure 3.2(a) we see that the dispersion of the predicted values is less than that of the observed values. Among the predicted values from different methods, those from Stone's procedure have the smallest dispersion, followed by those from the Bayesian alternative (1) approach. A careful examination of the plot indicates that the predictions of Stone's procedure are pulled toward the center of the data. In other words, Stone's procedure shrinks the prediction towards the center of the data. This phenomenon is strong in other clusters. Also, in other clusters, the Bayesian alternative (1) approach reveals the same shrinkage behavior. On the other hand the predicted values from the ordinary regression method have the widest dispersion and sometimes a dispersion wider than the observed values. This might be caused by outliers.

From Figure 3.2(b), the boxplots for prediction errors, we see that the prediction errors from the ordinary regression method have the widest dispersion, while the other

three methods produce prediction errors with almost the same dispersion. But in other clusters there are some features which are not present in the cluster we have highlighted in our discussion. In particular, the dispersions of the prediction errors from the Bayesian alternative (1) are sometimes smaller than those from Stone's procedure. In general, however, the Bayesian alternative (1) approach and Stone's procedure compete for the distinction of having prediction errors with smallest dispersion.

Figure 3.1(a): Relative Measure of Agreement for Sulfate

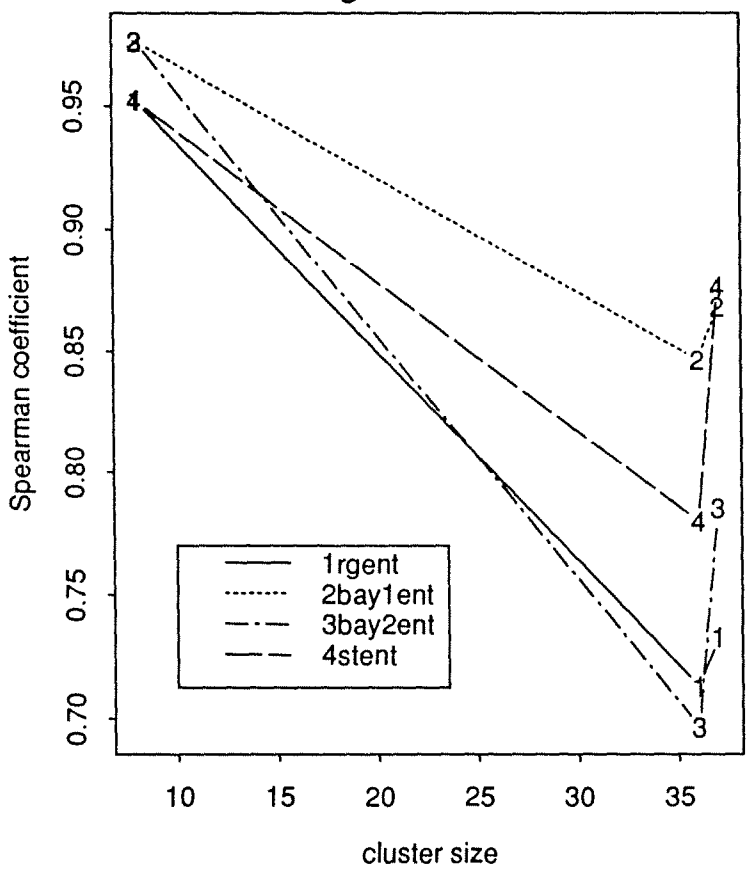
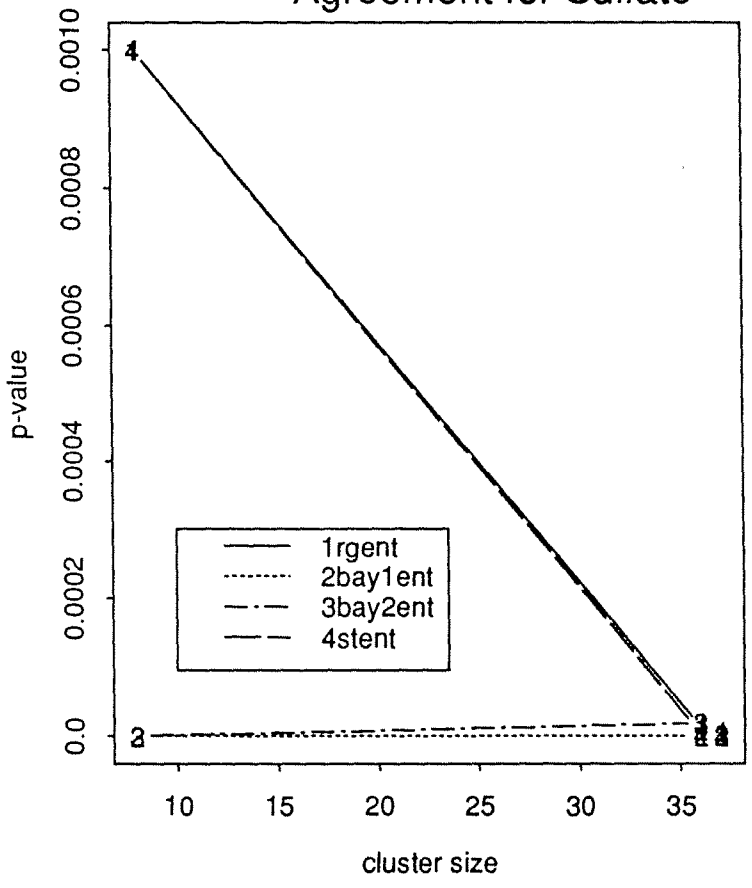


Figure 3.1(b): P-value for Tests of Agreement for Sulfate



Legend:
rgent = regression with entropy, bay1ent= Bayesian alternative (1) with entropy
bay1ent= Bayesian alternative (2) with entropy, stent = Stone's procedure with entropy

CHAPTER 4

DISCUSSION AND CONCLUSION

In the analysis of the monthly ion concentrations of stations in the NADP/NTN, we attempted to determine which station's ion concentrations were most accurately predicted from the other stations' concentrations. However, our analyses did not answer the question completely, due to three things. First, the predictability of stations were ranked within cluster, rather than across the entire network. Secondly, our analyses were conducted ion by ion. Thirdly, our different methods of prediction (ordinary regression, Bayes regression, and Stone's cross-validatory regression) resulted in different rankings. It may not be possible to completely resolve these three issues with this data set. However, we have gotten a clear picture of the relationship between the entropy approach and our prediction methods.

For each ion the analysis was done in clusters, each of size less than 48. This analysis by cluster makes the comparison of stations in different clusters impossible. But we argue that, since ion concentrations in different clusters are statistically different, we would not expect stations in different clusters to do well in predicting each other. However we cannot determine a single worst station from the network, since each cluster gives its own worst station. We suggest a tentative solution to this problem: to take the worst station from a larger cluster to be the overall network worst station. Since a station to be dropped out should be redundant, it seems logical to think of a redundant station coming from a larger cluster. Alternatively one might use geographical knowledge to make the decision. If one could accurately estimate missing values in the original data set, then one could use the weekly average concentrations with missing values replaced by their estimates. This would give us more than the 48 months as replicates, allowing analysis of the entire network, and hence give only one worst station for each ion. We

think this idea is feasible since Sten, Shen and Styr (1992) have used multiple regression to impute missing daily sulfate concentrations. The same problem would arise if we would want the overall network best station. We do not consider this problem to be of as much concern as the first one, since the need to retain only one station in a network is not realistic. But if there are reasons to identify the overall network best station, then other issues like geographical positions might be invoked in selecting the best station.

We did not carry out a formal analysis to compare each station's ranking in different ions, so we cannot say that, for instance, all of a particular station's ion concentrations are difficult to predict. So we cannot give a strong statement, but we have traced worst and best stations for all sulfate clusters and have given their ranks for each ion using the different prediction methods. Table 4.1 contains the ranks of the selected stations for the different ions. Since the clustering depends upon the ion under study, when looking at the rank of a station, we should note the cluster size to judge whether a station is towards the worst or the best position. From this table we see that the worst station as given by one ion is not necessarily the worst for another ion, and similarly for the best station. But some of the stations are put in the same category in either two of the ions or in all three ions. By category, we mean a ranking towards either the worst or best position. For example the station with identification code 025a is towards the worst position category in sulfate and nitrate while the station with identification code 163a is in the same category (towards the worst) in all three ions. Wu and Zidek (1992) found ion-to-ion differences in station clustering and station ranking. They pointed out that those differences might be informative and so resisted the use of a multivariate analysis at this stage. However, a multivariate analysis might give some indication of the simultaneous predictability of all of a station's ion concentrations.

Table 4.1: The Ranks of the Selected Stations in Different Ions.

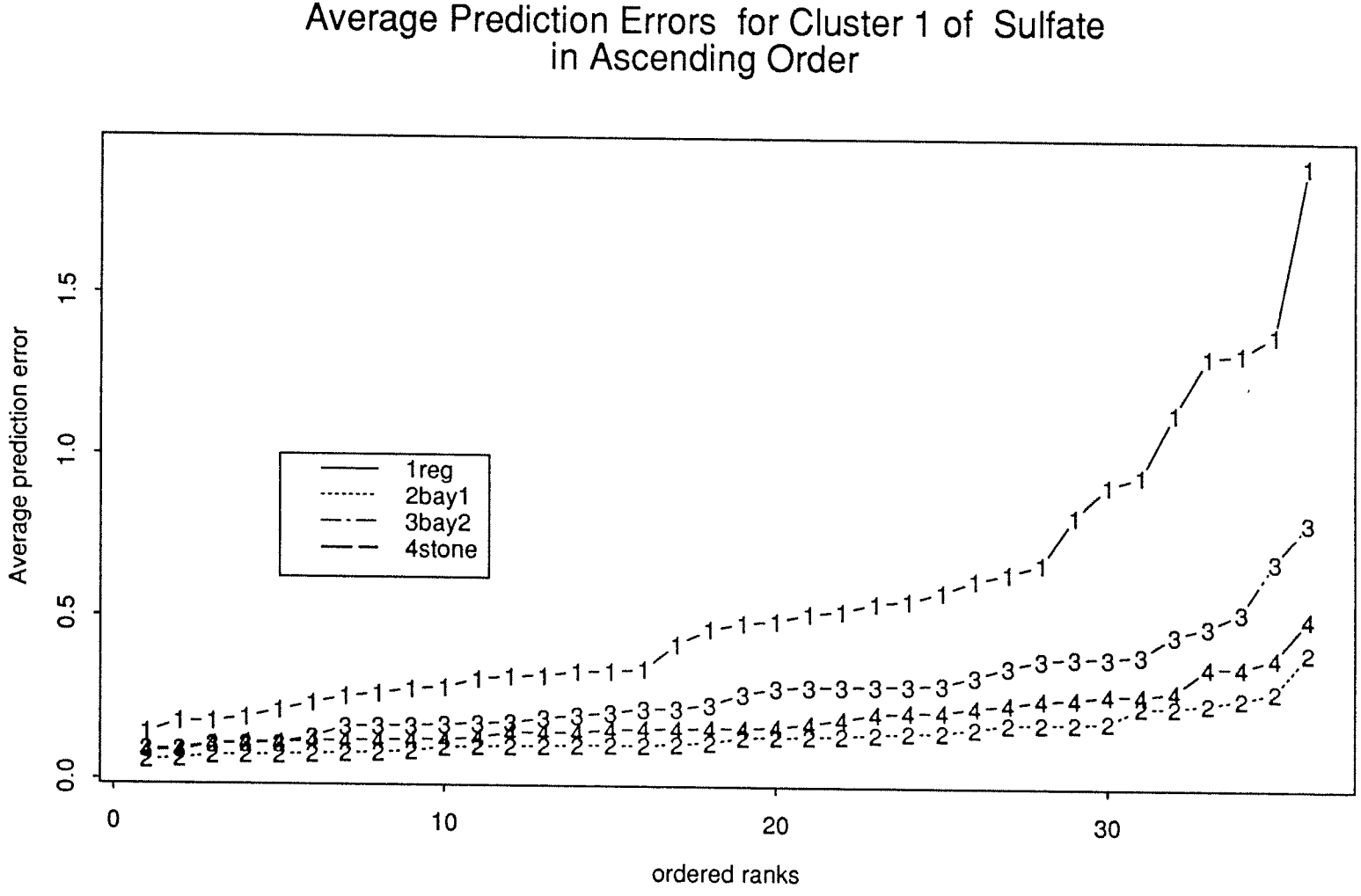
ion	method	station identification code					
		025a	249a	068a	281a	163a	037a
sulfate	regression	1	36	37	8	1	1
	Stone	2	36	36	7	1	1
	Bayesian 1	1	36	29	7	5	1
	Bayesian 2	1	36	37	7	1	1
	cluster size	36	36	37	8	37	8
hydrogen	regression	27	33	18	20	6	1
	Stone	32	26	19	5	2	4
	Bayesian 1	31	24	19	6	2	4
	Bayesian 2	32	24	19	6	2	4
	cluster size	33	33	20	20	18	20
nitrate	regression	10	7	20	30	11	27
	Stone	2	28	20	31	4	19
	Bayesian 1	8	14	20	30	6	16
	Bayesian 2	7	4	19	31	5	26
	cluster size	41	41	31	31	31	31

For the analysis of each ion within a cluster the different prediction methods' rankings did not always agree. In Chapter 3 we discussed the agreement of the station rankings from different methods in each cluster for each ion. It was found that the degree of agreement is reasonably high. We are interested in knowing which of the three ranking methods used in our study agrees strongly with those from the entropy based analysis of Wu and Zidek (1992). We found in Chapter 3 that the ranks from regression using a Bayesian alternative (1) approach agree with those from the entropy based approach

more often than the others. This implies that if, by using the entropy approach, a station found to be producing data of low quality is terminated, then in the future, regression using the Bayesian alternative (1) approach can be used to predict ion concentrations of the deleted station.

The close agreement of the ranking based on entropy analysis and the ranking from the methods used in our study reflects the relationship between the degree of the station's predictability and the amount of uncertainty reduced by the inclusion of a station into the network. That is, a station which is easily predicted will generally not noticeably reduce uncertainty when added into the network. Such a station is considered as producing data of low quality. On the other hand a station which is hard to predict will generally greatly reduce uncertainty when added into the network. Such a station is considered as producing data of high quality.

This exercise of ranking stations, starting with the worst station to the best station does, not mean that there is a worst station in an absolute sense. When we look at the trend of the average squared prediction errors (APE), we see that there is not much difference among the APE's for the first few worst stations. But the difference becomes sharper as we move towards the best station. Figure 4.1, a scatterplot of the average prediction errors (of cluster 1 of sulfate) against the ordered rankings, supports this fact. Another fact revealed by this figure is the close agreement of the average prediction errors from Stone's procedure and the regression using a Bayesian alternative (1) approach. These two methods have low average prediction errors when compared with the other methods. Scatterplots for other clusters for all three ions are included in the Appendix. This ranking exercise, as pointed out by Wu and Zidek (1992), can only be used to suggest the station which could be closed if there were budgetary constraints.


Average Prediction Errors for Cluster 1 of Sulfate
in Ascending Order
Average prediction error
0.0
0.5
1.0
1.5
0
10
20
30
ordered ranks
1reg
2bay1
3bay2
4stone
Legend:
reg=regression,bay1= Bayesian alternative (1), bay2= Bayesian alternative (2)
Stone = Stone's procedure

References

[1] Afifi, A.A and Elashoff R.M. (1961). Missing Observations in Multivariate Statistics, *Journal of the American Statistical Association*, **61**, 595-604.

[2] Buck, S.F. (1960). A Method of Estimation of Missing Values in Multivariate Data Suitable for Use with an Electronic Computer *Journal of the Royal Statiatical Society*, B, **22**, 302-307.

[3] Caselton, W.F., Kan. L. and Zidek, J.V. (1990). Quality Data Network Designs Based on Entropy. Unpublished Manuscript.

[4] Caselton, W.F., and Zidek, J.V. (1984). Optimal Monitoring Network Designs, *Statistics and Probability Letters*, **2**, 223-227.

[5] De Finetti, B. (1964). Foresight: its Logical Laws, its Subjective Sources. In *Studies in Subjective Probability* (H.E. Kyburgs Jr, and H.E. Smoker, Eds) pp 93-158. New York: Wiley.

[6] Frane J.W. (1976). Missing Data and BMDP: Some Pragmatic Approaches, *Technical Report*, No 45 Department of Biomathematics UCLA.

[7] Gibbons, J.D. (1976). Nonparametric Methods for Quantitative Analysis, Holt Rinehart and Winston.

[8] Hartigan, J.A. and Wong, M.A. (1979). A K-Means Clustering Algorithm, *Applied Statistics*, **28**, 101-108.

[9] Hewitt, E. and Savage, L.G. (1955). Symmetric Measures on Cartesian Products. *Trans. Amer. Math. Soc.*, **80**, 470-501.

[10] Krzanowski, W.J. and Lai Y.T. (1988). A Criterion for Determining the Number of Groups in a Data set Using Sum-of-Squares Clustering. *Biometrics.* **44**, 23-34.

[11] Lindley, D.V. and Smith, F.M. (1972). Bayes Esimation for Linear Model (with discussion). *Journal of the Royal Statistical Society*, B, **34**, 1-41.

[12] National Acid Precipitation Assessment Program, 1990. Integrated Assessment Report.

[13] Olsen, A.R. and Slavich A.l. (1986). Acid Precipitation in North America: 1984 Annual Data Summary fron Acid Deposition System Data Base. *U.S Environmental Protection Agency* , Research Triangle Park, NC. EPL/600/4-86/033.

[14] Sten, M.L., Shen, X and Styer, P.E. (1991). Applications of a Simple Regression Model to Acid Rain Data. Technical Report No 276. Department of Statistics. The University of Chicago, Chicago, IL.

[15] Stone, M. (1973). Cross-Validatory Choice and Assessment of Statistical Prediction *Journal of the Royal Statistical Society*, B, **36**, 111-132.

[16] Styer, P.E., and Stein, M.L. (1991). Acid Deposition Models for Detecting the Effects of Changes in Emissions, Technical Report No 331. Department of Statistics, University of Chicago, Chicago, IL.

[17] Wu, S. and Zidek, J.V. (1992) An Entropy Based Analysis of Data from Selected NAP/NTN Network Sites for 1983-86. Unpublished Manuscript.

APPENDIX

Table A1.1(a) Average Squared Prediction Errors and Ranks for the Stations in Cluster 1 of Hydrogen

station	regression		Bayesian 1		Bayesian 2		Stone		entropy
code	APE	rank	APE	rank	APE	rank	APE	rank	rank
015a	1.9848	16	1.0338	15	1.1952	15	1.0514	15	15
017a	0.6188	8	0.3383	6	0.3962	7	0.3787	7	5
024a	2.23042	17	1.2851	17	1.4310	17	1.2975	17	17
030a	0.37024	3	0.2929	4	0.3007	4	0.3143	4	7
036a	0.5956	7	0.4432	8	0.4952	9	0.4828	9	8
049a	0.3846	4	0.2732	3	0.2858	3	0.2965	3	3
051a	1.1721	13	0.6348	13	8179	13	0.7802	13	13
052a	0.5551	6	0.4543	9	0.4610	8	0.4527	8	10
053a	0.7862	11	0.5383	10	0.5945	11	0.6120	10	11
076a	0.9970	12	0.5514	11	0.7170	12	0.6206	11	9
077a	1.7739	15	1.1356	16	1.2522	16	1.1314	16	16
163a	0.3159	1	0.2261	2	2425	2	0.2370	2	2
252a	0.4828	5	0.3616	7	0.3893	6	0.3747	6	6
253a	0.3353	2	0.2160	1	0.2363	1	0.1979	1	1
258a	0.6336	9	0.3220	5	0.3720	5	0.3331	5	4
268a	1.3925	14	0.7729	14	0.9564	14	0.9725	14	14
283a	0.7388	10	0.5531	12	0.5669	10	0.6377	12	12
339a	2.5667	18	1.9549	18	1.8939	18	1.6876	18	18

Table A1.1(b) Average Squared Prediction Errors and Ranks for the Stations in Cluster 2 of Hydrogen

station	regression		Bayesian 1		Bayesian 2		Stone		entropy
code	APE	rank	APE	rank	APE	rank	APE	rank	rank
010a	1.0896	15	0.6209	13	o.7394	13	0.8512	14	14
011a	0.9865	12	0.5561	12	0.6864	12	0.7569	13	10
012a	1.7235	17	1.0377	16	1.3725	17	1.2563	17	16
016a	1.0326	14	0.7130	15	0.7983	15	0.7188	12	13
037a	0.3457	5	0.2518	4	0.2723	4	0.2668	4	4
038a	2.3707	19	1.4686	18	1.8701	18	1.4813	18	18
059a	0.6713	8	0.3935	7	0.4654	7	0.3520	7	6
061a	0.1679	1	0.1211	1	0.1266	1	0.1220	2	1
062a	3.4879	20	1.6470	20	2.2785	20	1.9624	20	20
068a	2.2189	18	1.6361	19	1.9070	19	1.6308	19	19
074a	0.3380	4	0.2929	5	0.2939	5	0.2885	6	8
078a	1.0004	13	0.5332	11	0.6751	11	0.6500	11	11
172a	0.1815	2	0.1264	2	0.1282	2	0.1001	1	2
173a	0.7678	10	0.4809	9	0.5937	10	0.5093	9	9
255a	1.3043	16	1.0602	17	1.1346	16	1.0255	16	17
271a	0.1905	3	0.1457	3	0.1520	3	0.1532	3	3
279a	0.7196	9	0.4438	8	0.5664	9	0.6342	10	7
280a	0.8923	11	0.6490	14	0.7683	14	0.9509	15	15
281a	0.3760	6	0.3082	6	0.3118	6	0.2765	5	5
354a	0.5384	7	0.5012	10	0.4830	8	0.4112	8	12

Table A1.1(c) Average Squared Prediction Errors and Ranks for the Stations in Cluster 3 of Hydrogen

station	regression		Bayesian 1		Bayesian 2		Stone		entropy
code	APE	rank	APE	rank	APE	rank	APE	rank	rank
007a	1.1711	1	1.1696	2	1.1837	2	1.1278	2	1
035a	2.0774	3	1.5934	3	1.5712	3	1.8186	3	3
070a	1.2163	2	1.1539	1	1.1760	1	1.1095	1	2

Table A1.1(d) Average Squared Prediction Errors and Ranks for the Stations in Cluster 4 of Hydrogen

station	regression		Bayesian 1		Bayesian 2		Stone		entropy
code	APE	rank	APE	rank	APE	rank	APE	rank	rank
004a	0.5734	1	0.4583	1	0.4688	1	0.5149	1	1
029a	3.5054	10	2.2386	10	2.6179	10	2.3558	9	9
034a	0.8194	6	0.7349	8	0.7374	7	0.7552	7	8
071a	1.7075	9	1.4480	9	1.5129	9	2.4643	10	10
166a	0.7058	4	0.6485	5	0.6552	5	0.5889	2	2
254a	0.8472	7	0.7255	7	0.7419	8	0.7550	6	4
273a	0.7375	5	0.6436	4	0.6527	4	0.6293	3	6
275a	0.7041	3	0.6008	2	0.6105	2	0.6944	5	3
282a	0.6755	2	0.6275	3	0.6359	3	0.6629	4	7
349a	0.8792	8	0.6991	6	0.7233	6	0.8074	8	5

Table A1.1(e) Average Squared Prediction Errors and Ranks for the Stations in Cluster 5 of Hydrogen

station	regression		Bayesian 1		Bayesian 2		Stone		entropy
code	APE	rank	APE	rank	APE	rank	APE	rank	rank
020a	0.4309	15	0.1250	10	0.1601	7	0.1851	12	12
021a	1.2228	30	0.6873	33	0.7766	33	0.5594	31	31
022a	0.5742	19	0.2189	16	0.3526	14	0.2292	16	22
023a	0.3755	13	0.1311	12	0.1838	10	0.1721	9	9
025a	0.9739	27	0.5283	31	0.7410	32	0.6173	32	26
028a	1.1945	29	0.4046	28	0.6830	27	0.5428	29	32
031a	0.7658	24	0.1811	14	0.3620	15	0.1566	7	20
032a	0.5223	18	0.1411	13	0.1724	8	0.1700	8	13
033a	0.6182	21	0.1816	15	0.3987	18	0.1735	10	16
039a	1.7682	32	0.2804	23	0.4872	20	0.4494	28	30
040a	0.3632	11	0.1001	7	0.1820	9	0.2355	17	10
041a	1.0391	28	0.2714	22	0.4540	19	0.28484	21	23
046a	0.2554	5	0.0919	5	0.1358	5	0.1256	4	2
047a	0.3441	10	0.1139	8	0.1947	11	0.17474	11	6
055a	0.2656	6	0.0948	6	0.2010	12	0.0984	3	7
056a	0.1266	2	0.0892	4	0.1034	3	0.0972	2	3
058a	0.4510	16	0.0849	3	0.1181	4	0.1857	13	4
063a	0.3636	12	0.1302	11	0.2541	13	0.2176	15	14
064a	0.1625	3	0.0806	2	0.0923	2	0.1259	5	8

Table A1.1(e) Continued

station	regression		Bayesian 1		Bayesian 2		Stone		entropy
code	APE	rank	APE	rank	APE	rank	APE	rank	rank
065b	0.1201	1	0.0645	1	0.0905	1	0.0961	1	1
073a	1.7031	31	0.2712	21	0.7115	29	0.5487	30	27
075a	0.3038	8	0.1238	9	0.1473	6	0.1546	6	11
161a	0.6085	20	0.5617	32	0.6372	25	0.7598	33	33
164a	0.7928	25	0.2387	18	0.3891	16	0.4198	27	24
168a	0.3324	9	0.4531	29	0.6983	28	0.3153	22	19
171a	0.6740	23	0.5213	30	0.6705	26	0.2420	18	17
249a	2.2977	33	0.2988	24	0.5887	24	0.4195	26	22
251a	0.4256	14	0.3009	25	0.5016	22	0.3832	25	28
257a	0.9505	26	0.2470	20	0.4876	21	0.3260	24	18
272a	0.6684	22	0.2424	19	0.5166	23	0.2503	20	21
277a	0.2901	7	0.3835	27	0.7221	30	0.1886	14	5
285a	0.2542	4	0.2363	17	0.3905	17	0.2461	19	15
350a	0.4782	17	0.3247	26	0.7305	31	0.3239	23	25

Table A1.1(f) Average Squared Prediction Errors and Ranks for the Stations in Cluster 6 of Hydrogen

station	regression		Bayesian1		Stone		entropy
code	APE	rank	APE	rank	APE	rank	rank
160a	1.5016	1	1.5016	1	1.4745	1	1
278a	1.9126	2	1.9125	2	1.8781	2	2

Table A1.2(a): Average Squared prediction Errors and Ranks for the Stations in Cluster 1 of Sulfate

station	regression		Bayesian 1		Bayesian 2		Stone		entropy
code	APE	rank	APE	rank	APE	rank	APE	rank	rank
004a	0.4139	6	0.1514	3	0.2577	7	0.1512	8	5
010a	0.5927	12	0.3185	25	0.3822	14	0.2303	15	20
011a	0.8626	21	0.2037	16	0.5076	21	0.4108	30	16
012a	0.9378	24	0.4955	34	0.6559	27	0.4509	33	33
029a	0.6013	13	0.2853	22	0.4086	16	0.2763	20	22
030a	0.4442	7	0.3079	24	0.3234	11	0.2680	18	24
034a	0.4524	8	0.1286	2	0.2102	3	0.1356	4	7
035a	0.7034	16	0. 3696	30	0.4618	18	0.3128	23	25
036a	0.5708	10	0.1943	13	0.3089	9	0.1424	6	9
038a	0.6547	15	0.1985	15	0.3573	13	0.2309	16	17
039a	0.9386	25	0.4242	32	0.6640	28	0.4877	35	36
052a	0.7357	17	0.2816	21	0.4071	15	0.3000	22	21
068a	3.1738	37	0.3626	29	1.5558	37	0.5189	36	31
070a	1.7942	35	0.4356	33	0.9571	35	0.4019	29	35
071a	0.8449	20	0.3428	28	0.6376	25	0.3743	28	32
076a	0.9870	26	0.1917	12	0.5618	24	0.2728	19	14
077a	1.1916	31	0.3854	31	0.6642	29	0.3134	31	27
160a	1.2444	32	0.6412	35	0.7963	33	0.7200	37	37
163a	0.1773	1	0.1577	5	0.1482	1	0.1185	1	10

Table A1.2(a) Continued

station	regression		Bayesian 1		Bayesian 2		Stone		entropy
code	APE	rank	APE	rank	APE	rank	APE	rank	rank
164a	0.3223	5	0.1818	11	0.2514	6	0.2022	10	13
166a	0.6382	14	0.2902	23	0.4827	20	0.2271	14	8
172a	0.7576	18	0.2284	17	0.3448	12	0.2814	21	29
173a	0.3154	3	0.1681	7	0.2182	4	0.1789	9	2
252a	0.3208	4	0. 1746	9	0.2260	5	0.1285	2	6
253a	0.8422	19	0.1529	4	0.4661	19	0.1419	5	11
254a	1.3770	34	0.2710	20	0.6783	30	0.3467	26	23
255a	1.1435	30	0.6489	37	0.7348	31	0.4369	32	30
257a	0.5785	11	0.1604	6	0.3167	10	0.2579	17	4
258a	0.1.0605	27	0.1800	8	0.5520	22	0.2261	12	18
268a	1.1218	29	0.6447	36	0.9270	34	0.4865	34	34
273a	0.2449	2	0.1973	14	0.1844	2	0.1504	7	3
275a	0.8560	22	0.2527	18	0.6461	26	0.2268	13	12
278	1.3028	33	0.3208	26	0.7546	32	0.3410	24	28
279a	2.3086	36	0.2603	19	1.1182	36	0.3448	25	26
280a	0.9371	23	0.3369	27	0.5617	23	0.3590	27	19
282a	1.1098	28	0.1778	10	0.4526	17	0.2030	11	15
349a	0.4750	9	0.1192	1	0.2893	8	0.1321	3	1

Table A1.2(b): Average Squared Prediction Errors and Ranks for the Stations in Cluster 2 of Sulfate

station	regression		Bayesian 1		Bayesian 2		Stone		entropy
code	APE	rank	APE	rank	APE	rank	APE	rank	rank
017a	1.3291	34	0.1449	23	0.5271	34	0.2528	28	29
020a	0.1888	4	0.0608	2	0.1078	4	0.0843	1	3
021a	0.3333	14	0.1646	26	0.1762	12	0.1181	7	21
022a	0.5234	22	0.1127	17	0.2933	23	0.1222	9	19
023a	0.4187	17	0.1028	12	0.2672	19	0.1168	6	14
024a	0.6716	28	0.1069	14	2899	21	0.1224	10	17
025a	0.1418	1	0.0576	1	0.0901	1	0.0849	2	4
028a	0.3180	12	0.2334	32	0.2201	16	0.2571	29	34
031a	1.3176	33	0.1814	29	0.4555	32	0.2687	31	28
032a	0.6416	27	0.1403	22	0.2983	25	0.1907	22	23
033a	0.4842	19	0.0987	10	0.2341	18	0.1457	13	9
040a	0.4650	18	0.0824	8	0.2211	17	0.2070	23	32
041a	0.3411	16	0.1044	13	0.1637	8	0.1551	15	12
046a	0.1753	2	0.0724	4	0.1288	6	0.1203	8	5
047a	0.2546	7	0.1173	18	0.1725	11	0.1658	20	13
049a	1.1425	32	0.2818	35	0.3762	28	0.3556	33	33
051a	0.6200	26	0.1074	15	0.3843	30	0.2287	26	15
053a	1.3871	35	0.2607	34	0.6844	35	0.2801	32	31

Table A1.2(b) Continued

station	regression		Bayesian 1		Bayesian 2		Stone		entropy
code	APE	rank	APE	rank	APE	rank	APE	rank	rank
055a	0.1759	3	0.0710	3	0.1091	5	0.1058	3	2
056a	0.3197	13	0.0774	6	0.2069	15	0.1463	14	6
058a	0.5563	24	0.1076	16	0.3542	27	0.1588	17	11
063a	0.2831	10	0.1369	21	0.1696	10	0.2126	24	16
064a	0.2331	6	0.0800	7	0.0904	2	0.1114	5	20
065b	0.2114	5	0.0877	9	0.1069	3	0.1113	4	1
073a	0.4927	20	0.1788	27	0.3248	26	0.3566	34	26
075a	0.3340	15	0.0737	5	0.1616	7	0.1614	19	7
161a	0.3083	11	0.1519	25	0.1645	9	0.1581	16	27
168a	0.9466	31	0.2448	33	0.4810	33	0.3853	35	35
171a	0.5840	25	0.1354	20	0.3917	31	0.1766	21	22
249a	1.9059	36	0.4026	36	0.8033	36	0.5058	36	36
251a	0.9152	30	0.1861	30	0.2904	22	0.2146	25	30
272a	0.2783	9	0.1025	11	0.1894	13	0.1289	11	8
277a	0.2635	8	0.1335	19	0.1986	14	0.1449	12	10
283a	0.5486	23	0.2327	31	0.3832	29	0.1608	18	25
285a	0.5149	21	0.1798	28	0.2947	24	0.2379	27	24
350a	0.8198	29	0.1493	24	0.2854	20	0.2680	30	18

Table A1.2(c): Average Squared Prediction Errors and Ranks for the Stations in Cluster 3 of Sulfate

station	Regression		Bayesian 1		Bayesian 2		Stone		entropy
code	APE	rank	APE	rank	APE	rank	APE	rank	rank
037a	0.1484	1	0.1343	1	0.1334	1	0.1483	1	1
059a	0.3091	5	0.2683	5	0.2711	5	0.2836	5	4
061a	0.2537	4	0.2354	4	0.2403	4	0.2279	4	5
074a	0.4518	6	0.3693	6	0.3812	6	0.4096	6	6
078a	0.1796	2	0.1635	2	0.1659	2	0.1610	2	2
271a	0.2498	3	0.1979	3	0.1983	3	0.2276	3	3
281a	0.5258	8	0.4044	7	0.4604	7	0.5398	8	7
354a	0.5053	7	0.4773	8	0.4767	8	0.4472	7	8

Table A1.3(a): Average Squared Prediction Errors and Ranks for the Stations in Cluster 1 of Nitrate

station	regression		Bayesian 1		Bayesian 2		Stone		entropy
code	APE	rank	APE	rank	APE	rank	APE	rank	rank
004a	3.044	40	0.1882	9	0.1.2349	38	0.182	15	29
011a	1.999	33	0.3615	26	0.6625	28	0.217	24	32
012a	0.752	14	0.6384	39	0.6911	29	0.326	36	39
020a	2.109	34	0.2576	18	0.7614	34	0.157	8	3
021a	0.999	20	0.2444	15	0.7079	30	0.261	30	17
022a	1.815	30	0.4309	32	0.5707	24	0.210	20	11
023a	0.660	11	0.2570	17	0.3090	12	0.161	9	7
024a	0.782	15	0.4638	34	0.4864	17	0.298	33	37
025a	0.613	10	0.1856	8	0.2672	7	0.095	2	2
031a	1.738	29	0.3347	24	0.4893	18	0.192	17	30
032a	0.461	8	0.2451	16	0.2904	9	0.235	27	20
033a	0.555	9	0.1725	7	0.3006	10	0. 155	6	12
038a	1.471	26	0.6027	37	0.5710	25	0.217	23	21
040a	2.632	37	0.1124	3	0.7176	31	0.162	10	33
041a	0.701	13	0.3007	21	0.3036	11	0.211	21	18
046a	1.522	27	0.1496	4	0.3389	13	0.132	3	5
047a	0.850	17	0.2017	11	0.5010	21	0.171	11	6
051a	1.823	31	0.5791	36	0.7359	33	0.296	32	22

Table A1.3(a) Continued

station	regression		Bayesian 1		Bayesian 2		Stone		entropy
code	APE	rank	APE	rank	APE	rank	APE	rank	rank
053a	1.173	21	0.4834	35	0.8777	35	0.417	39	40
055a	0.393	4	0.1682	5	0.2056	5	0.151	5	4
056a	0.952	19	0.1925	10	0.4903	19	0.178	13	10
058a	0.401	5	0.2761	20	0.2590	6	0.172	12	8
063a	0.246	3	0.2021	12	0.1865	3	0.156	7	14
064a	1.462	25	0.1712	6	0.5427	23	0.186	16	13
065a	0.227	2	0.1086	2	0.1122	2	0.134	4	1
076a	0.860	18	0.3605	25	0.4248	16	0.210	19	15
077a	2.716	38	0.7553	40	1.8189	40	0.343	37	38
161a	0.435	6	0.2675	19	0.2726	8	0.234	26	25
166a	0.844	16	0.3775	29	0.5725	26	0.219	25	23
168a	2.435	36	0.4186	31	0.8832	36	0.314	35	31
171a	2.760	39	0.3222	22	0.6364	27	0.201	18	27
249a	0.436	7	0.2332	14	0.2053	4	0.241	28	35
251a	1.265	23	0.3682	27	0.5335	22	0.472	40	36
252a	1.705	28	0.6175	38	0.7186	32	0.284	31	28
253a	2.374	35	0.3954	30	0.9132	37	0.211	22	24
258a	1.263	22	0.2126	13	0.3455	14	0.181	14	16

Table A1.3(a) Continued

station	regression		Bayesian 1		Bayesian 2		Stone		entropy
code	APE	rank	APE	rank	APE	rank	APE	rank	rank
272a	1.437	24	0.3252	23	0.4922	20	0.255	29	19
273a	1.920	32	0.8116	41	1.4461	39	0.298	34	34
277a	3.651	41	0.3754	28	1.8964	41	0.512	41	41
283a	0.106	1	0.0525	1	0.0539	1	0.094	1	9
285a	0.684	12	0.4368	33	0.3922	15	0.370	38	26

Table A1.3(b) Average Squared Prediction Error and Ranks for the Stations in Cluster 2 of Nitrate

station	regression		Bayesian 1		Bayesian 2		Stone		entropy
code	APE	rank	APE	rank	APE	rank	APE	rank	rank
010a	0.777	4	0.2596	8	0.4208	8	0.299	5	8
017a	0.341	1	0.2267	2	0.2502	1	0.235	1	3
028a	1.511	16	0.8503	26	1.0421	18	0.477	14	10
029a	0.791	5	0.2202	1	0.3050	4	0.465	13	2
030a	2.351	24	0.3922	13	0.9411	16	1.133	27	21
036a	1.201	12	0.3111	9	0.5561	10	0.368	7	6
037a	3.059	27	0.4473	16	1.5074	26	0.659	19	20
049a	1.400	14	0.9118	27	1.2118	25	0.745	21	27
052a	0.569	3	0.2420	5	0.2613	2	0.251	2	5
068a	1.910	20	0.5580	20	1.044	19	0.704	20	26
070a	2.112	22	0.4465	15	0.9874	17	0.394	9	24
071a	0.866	8	0.3748	11	0.4883	9	0.390	8	14

Table A1.3(b) Continued

station	regression		Bayesian 1		Bayesian 2		Stone		entropy
code	APE	rank	APE	rank	APE	rank	APE	rank	rank
073a	1.626	17	0.5946	21	0.8578	14	0.502	15	15
078a	1.421	15	0.4677	17	0.5850	12	0.644	18	13
160a	1.670	18	1.0074	29	1.1883	23	1.023	25	28
163a	1.164	11	0.2510	6	0.3503	5	0.280	4	11
164a	0.393	2	0.2272	3	0.2668	3	0.310	6	7
173a	1.104	9	0.2545	7	0.5617	11	0.403	10	1
254a	1.728	19	0.4967	18	0.8702	15	0.587	17	12
255a	2.831	26	0.3751	12	1.1103	21	0.412	11	16
257a	0.850	7	0.3524	10	0.4157	7	0.453	12	17
268a	2.190	23	1.9298	31	1.9351	29	1.628	30	29
271a	4.333	31	0.3952	14	1.6762	27	0.516	16	18
275a	1.338	13	0.6996	24	1.1334	22	1.077	26	22
278a	2.420	25	0.9904	28	1.2061	24	1.249	29	31
279a	3.426	28	0.7868	25	1.7865	28	1.209	28	19
280a	3.694	29	0.5129	19	2.1104	30	0.872	22	23
281a	4.231	30	1.0187	30	2.4678	31	1.850	31	30
282a	0.811	6	0.2357	4	0.0.4108	6	0.278	3	4
349a	1.921	21	0.6867	23	1.0605	20	0.887	23	9
354a	1.155	10	0.6291	22	0.7954	13	0.997	24	25

Table A1.3(c) Average Squared Prediction Errors and Ranks for the Stations in Cluster 3 of Nitrate

station	regression		Bayesian 1		Bayesian 2		Stone		entropy
code	APE	rank	APE	rank	APE	rank	APE	rank	rank
059a	0.982	4	0.9444	4	0.9478	4	1.005	4	4
061a	0.911	3	0.8603	3	0.8442	3	0.858	3	3
074a	0.653	2	0.6563	2	0.6572	2	0.737	2	2
172a	0.400	1	0.3996	1	0.4045	1	0.371	1	1

Table A2.1(a): Association Measures for Cluster 1 of Hydrogen

methods being compared	R or W	Q	p-value
regression versus Bayesian 1	0.952	-	0.00004
regression versus Bayesian 2	0.967	-	0.00003
regression versus Stone	0.961	-	0.00004
regression versus entropy	0.911	-	0.00008
Bayesian 1 versus Bayesian 2	0.989	-	0.00002
Bayesian 1 versus Stone	0.996	-	0.00002
Bayesian 1 versus entropy	0.981	-	0.00003
Bayesian 2 versus Stone	0.994	-	0.00002
Bayesian 2 versus entropy	0.967	-	0.00003
Stone versus entropy	0.975	-	0.00003
all methods	0.975	82.923	0.000

Table A2.1(b): Association Measures for Cluster 2 of Hydrogen

methods being compared	R or W	Q	p-value
regression versus Bayesian 1	0.973	-	0.00001
regression versus Bayesian 2	0.982	-	0.0.000009
regression versus Stone	0.970	-	0.00001
regression versus entropy	0.938	-	0.00002
Bayesian 1 versus Bayesian 2	0.994	-	0.000007

Table A2.1(b): Continued

methods being compared	R or W	Q	p-value
Bayesian 1 versus Stone	0.980	-	0.000009
Bayesian 1 versus entropy	0.980	-	0.000009
Bayesian 2 versus Stone	0.986	-	0.000008
Bayesian 2 versus entropy	0.967	-	0.00001
Stone versus entropy	0.967	-	0.00001
all methods	0.979	93.011	0.0000

Table A2.1(c): Association Measures for Cluster 3 of Hydrogen

methods being compared	R or W	S	p-value
regression versus Bayesian 1	0.50	-	0.500
regression versus Bayesian 2	0.50	-	0.500
regression versus Stone	0.50	-	0.500
regression versus entropy	1.00	-	0.167
Bayesian 1 versus Bayesian 2	1.00	-	0.167
Bayesian 1 versus Stone	1.000	-	0.167
Bayesian 1 versus entropy	0.50	-	0.500
Bayesian 2 versus Stone	1.00	-	0.167
Bayesian 2 versus entropy	0.50	-	0.500
Stone versus entropy	0.50	-	0.500
all methods	0.76	38	0.024

Table A2.1(d): Association Measures for Cluster 4 of Hydrogen

methods being compared	R or W	Q	p-value
regression versus Bayesian 1	0.927	-	0.000
regression versus Bayesian 2	0.939	-	0.000
regression versus Stone	0.879	-	0.001
regression versus entropy	0.673	-	0.019
Bayesian 1 versus Bayesian 2	0.988	-	0.000
Bayesian 1 versus Stone	0.830	-	0.002
Bayesian 1 versus entropy	0.745	-	0.009
Bayesian 2 versus Stone	0.818	-	0.003
Bayesian 2 versus entropy	0.697	-	0.015
Stone versus entropy	0.782	-	0.005
all methods	0.862	38.804	0.000012

Table A2.1(e): Association Measures for Cluster 5 of Hydrogen

methods being compared	R or W	Q	p-value
regression versus Bayesian 1	0.637	-	0.00016
regression versus Bayesian 2	0.618	-	0.00024
regression versus Stone	0.739	-	0.000015
regression versus entropy	0.810	-	0.000002
Bayesian 1 versus Bayesian 2	0.945	-	0.0000

Table A2.1(e): Continued

Bayesian 1 versus Stone	0.839	-	0.0000011
Bayesian 1 versus entropy	0.811	-	0.0000022
Bayesian 2 versus Stone	0.813	-	0.0000021
Bayesian 2 versus entropy	0.753	-	0.00001
Stone versus entropy	0.881	-	0.00000029
all methods	0.828	132.42	0.0000

Table A2.1(f): Association Measures for Cluster 6 of Hydrogen

methods being compared	R or W	S	p-value
regression versus Bayesian 1	1.0	-	0.5
regression versus Bayesian 2	1.0	-	0.5
regression versus Stone	1.0	-	0.5
regression versus entropy	1.0	-	0.5
Bayesian 1 versus Bayesian 2	1.0	-	0.5
Bayesian 1 versus Stone	1.0	-	0.5
Bayesian 1 versus entropy	1.0	-	0.5
Bayesian 2 versus Stone	1.0	-	0.5
Bayesian 2 versus entropy	1.0	-	0.5
Stone versus entropy	1.0	-	0.5
all methods	1.0	8	-

Table A2.2(a): Association Measures for Cluster 1 of Sulfate

methods being compared	R or W	Q	p-value
regression versus Bayesian 1	0.705	-	0.000015
regression versus Bayesian 2	0.919	-	0.000
regression versus Stone	0.752	-	0.000004
regression versus entropy	0.713	-	0.000012
Bayesian 1 versus Bayesian 2	0.726	-	0.0000086
Bayesian 1 versus Stone	0.793	-	0.0000014
Bayesian 1 versus entropy	0.847	-	0.0000002
Bayesian 2 versus Stone	0.755	-	0.0000039
Bayesian 2 versus entropy	0.697	-	0.000018
Stone versus entropy	0.780	-	0.0000019
all methods	0.815	142.61	0.0000

Table A2.2(b): Association Measures for Cluster 2 of Sulfate

methods being compared	R or W	S	p-value
regression versus Bayesian 1	0.973	-	0.000011
regression versus Bayesian 2	0.982	-	0.0000093
regression versus Stone	0.970	-	0.000012
regression versus entropy	0.938	-	0.000022
Bayesian 1 versus Bayesian 2	0.994	-	0.0000074

Table A2.2(b) : Continued

Bayesian 1 versus Stone	0.980	-	0.0000096
Bayesian 1 versus entropy	0.980	-	0.0000096
Bayesian 2 versus Stone	0.986	-	0.0000086
Bayesian 2 versus entropy	0.967	-	0.000012
Stone versus entropy	0.967	-	0.000012
all methods	0.979	93.011	0.0000

Table A2.2(c): Association Measures for Cluster 3 of Sulfate

methods being compared	R or W	Q	p-value
regression versus Bayesian 1	0.976	-	0.000
regression versus Bayesian 2	0.976	-	0.000
regression versus Stone	1.000	-	0.000
regression versus entropy	0.952	-	0.001
Bayesian 1 versus Bayesian 2	1.000	-	0.000
Bayesian 1 versus Stone	0.976	-	0.0000
Bayesian 1 versus entropy	0.976	-	0.000
Bayesian 2 versus Stone	0.976	-	0.000
Bayesian 2 versus entropy	0.976	-	0.000
Stone versus entropy	0.952	-	0.001
all methods	0.981	34.3333	0.000

Table A2.3(a): Association Measures for Cluster 1 of Nitrate

methods being compared	R or W	Q	p-value
regression versus Bayesian 1	0.394	-	0.0064
regression versus Bayesian 2	0.877	-	0.0000
regression versus Stone	0.082	-	0.3
regression versus entropy	0.455	-	0.002
Bayesian 1 versus Bayesian 2	0.587	-	0.00010
Bayesian 1 versus Stone	0.033	-	0.4
Bayesian 1 versus entropy	0.638	-	0.000027
Bayesian 2 versus Stone	0.108	-	0.2
Bayesian 2 versus entropy	0.572	-	0.00015
Stone versus entropy	0.241	-	0.06
all methods	0.514	102.73	0.0000

Table A2.3(b): Association Measures for Cluster 2 of Nitrate

methods being compared	R or W	Q	p-value
regression versus Bayesian 1	0.65	-	0.00018
regression versus Bayesian 2	0.897	-	0.00000047
regression versus Stone	0.0411	-	0.4
regression versus entropy	0.664	-	0.00014
Bayesian 1 versus Bayesian 2	0.829	-	0.0000029

Table A2.3(b): Continued

Bayesian 1 versus Stone	0.002	-	0.49
Bayesian 1 versus entropy	0.799	-	0.0000059
Bayesian 2 versus Stone	0.083	-	0.32
Bayesian 2 versus entropy	0.772	-	0.000012
Stone versus entropy	0.102	-	0.28
all methods	0.567	85.059	0.0000

Table A2.3(c): Association Measures for Cluster 3 of Nitrate

methods being compared	R or W	S	p-value
regression versus Bayesian 1	1.0	-	0.042
regression versus Bayesian 2	1.0	-	0.042
regression versus Stone	1.0	-	0.042
regression versus entropy	1.0	-	0.042
Bayesian 1 versus Bayesian 2	1.0	-	0.042
Bayesian 1 versus Stone	1.0	-	0.042
Bayesian 1 versus entropy	1.0	-	0.042
Bayesian 2 versus Stone	1.0	-	0.042
Bayesian 2 versus entropy	1.0	-	0.042
Stone versus entropy	1.0	-	0.042
all methods	1.0	125	0.0018

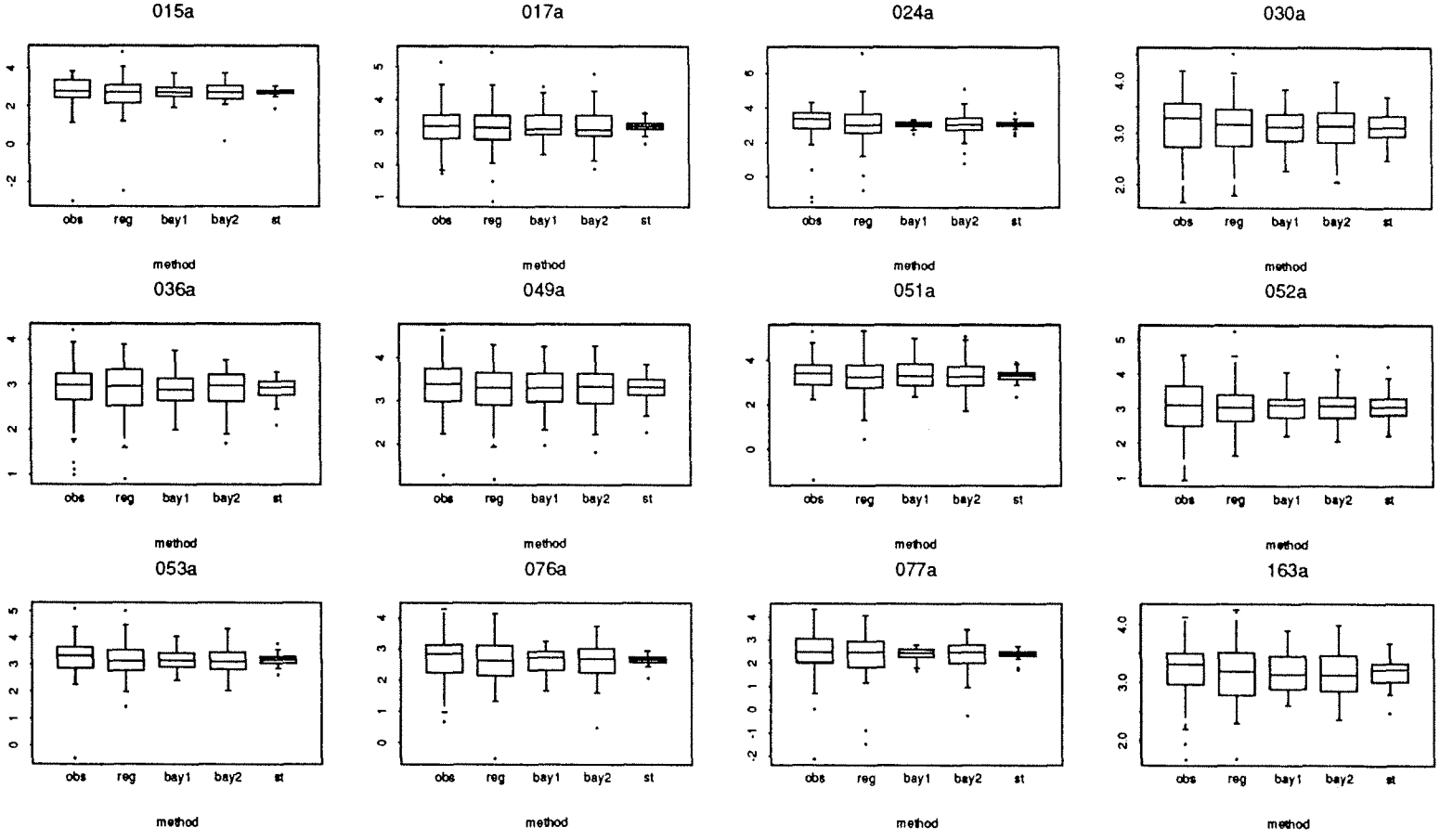
Table A3: Names and Identification Codes for the Sites Included in the Study

site ID	site name	site ID	site name
004a	Fayetteville, Arkansas	070a	K-Bar, Texas
007a	Hopland (Ukiah), California	071a	Victoria, Texas
010a	Rocky Mt. Net park, colorado	073a	Horton's Station, virginia
011a	Manitou, Colorado	074a	Olympic Nat.park, Washington
012a	Pawnee, Colorado	075a	Parsons, West Virginia
015a	Bradford Forest, Florida	076a	Trout Lake, Wisconsin
016a	Everglades Nat.Pa, Florida	077a	Spooner, Wisconsin
017a	Georgia Station, Georgia	078a	Yellowstone, Wyoming
020a	Bondville, Illinois	160a	Alamosa, Colorado
021a	Argonne, Illinois	161a	Salem, Illinois
022a	Southern Ill U, Illinois	163a	Caribou (a), Maine
023a	Dixon Springs Illinois	164a	Bridgton, Maine
024a	NIARC, Illinois	166a	Fernberg, Minnesota
025a	Idiaana Dunes, Indiana	168a	Huntington, New York
028a	Elkmont, Tennessee	171a	WalkerBranch, Tennessee
029a	Mesa Verde, Colorado	172a	American Samoa, American Samoa
030a	Greenville Station, Maine	173a	Sand Spring, Colorado
031a	Douglas Lake, Michigan	249a	Bennington, Vermont
032a	Kellogg, Micigan	251a	NACL, Massachusetts
033a	Wellston, Michigan	252a	Ashland, Missouri
034a	Marcell, Minnesota	253a	University Forest, Missouri
035a	Lamberton, Minnesota	254a	Forest Seed Ctr, Texas

Table A3: Continued

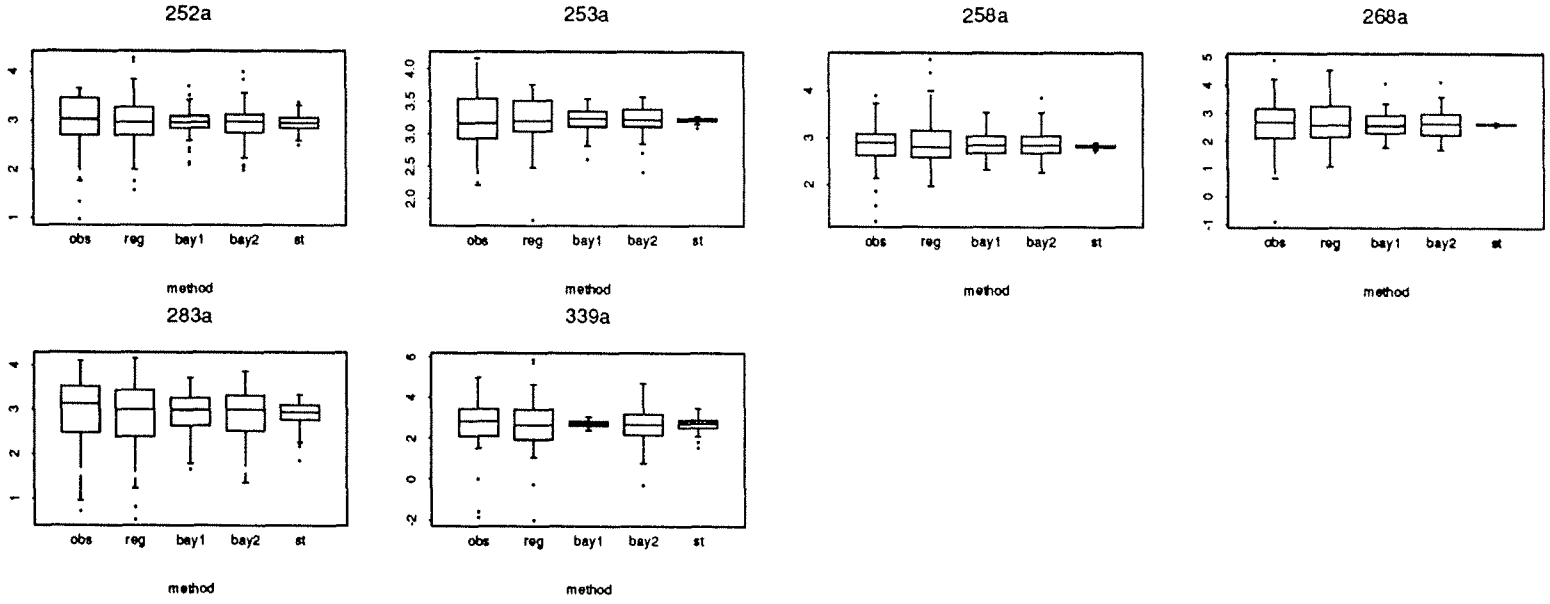
036a	Meridian, Mississippi	255a	Newcastie, Wyoming
037a	Glacier Nat. Park, Montana	257a	Acadia > 11/81, Maine
038a	Mead, Nebraska	258	Chassell, Michigan
039a	Hubbard Brook, New Hampshire	268a	Warren ZWSW, Arkansas
040a	Aurora, New York	271a	Headquarters, Idaho
041a	Chautauqua, New York	272a	Purdue U Ag Farm, Indiana
046a	Bennett Bridge, New York	273a	Konza Prairie, Kansas
047a	Jasper, New York	275	Iberia, Louisiana
049a	Lewiston, North Carolina	277a	East, Massachusetts
051a	Piedmont Station, North Carolina	278a	Give Out Morgan, Montana
052a	Clinton Station, North Carolina	279a	Bandelier, New Mexico
053a	Finley (a), North Calorina	280	Cuba, New Mexico
055a	Delaware, Ohio	281a	Bull Run, Oregon
056a	Caldwell, Ohio	282a	Longview, Texas
058a	Wooster, Ohio	283a	Lake Bubay, Wisconsin
059a	Alsea, Oregon	285a	Washington Xing, New Jersey
061a	H.J. Andrews, Oregon	339a	Bellville, Georgia
062a	Teddy Roosevelt NP, North Dakota	349a	Southeast, Louisiana
063a	Kane, Pennsylvania	350	Wye, Maryland
064a	Leading Ridge, Pennsylvania	354a	St. Mary Ranger St, Montana
065b	Penn State, Pennsylvania		
068a	Grand Canyon, Arizona		

Figure A1.1(a): Boxplots of Observed and Predicted Values for Cluster 1 of Hydrogen



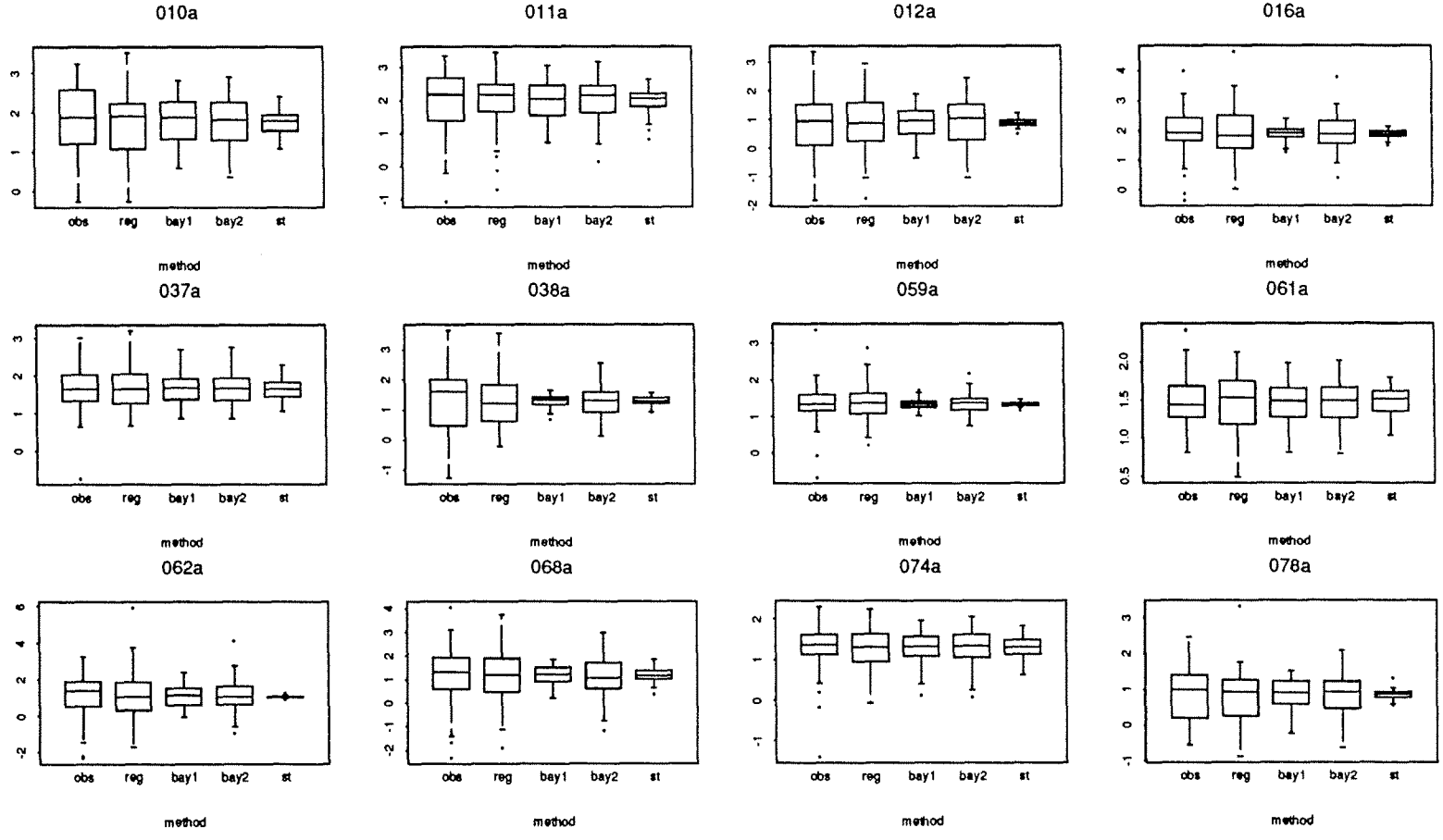
Legend:
obs=observed, reg=ordinary regression, bay1=regression using a Bayesian alternative (1) approach, bay2 = Bayesian alternative (2) approach, st= Stone's procedure

Figure A1.1(a): Continued



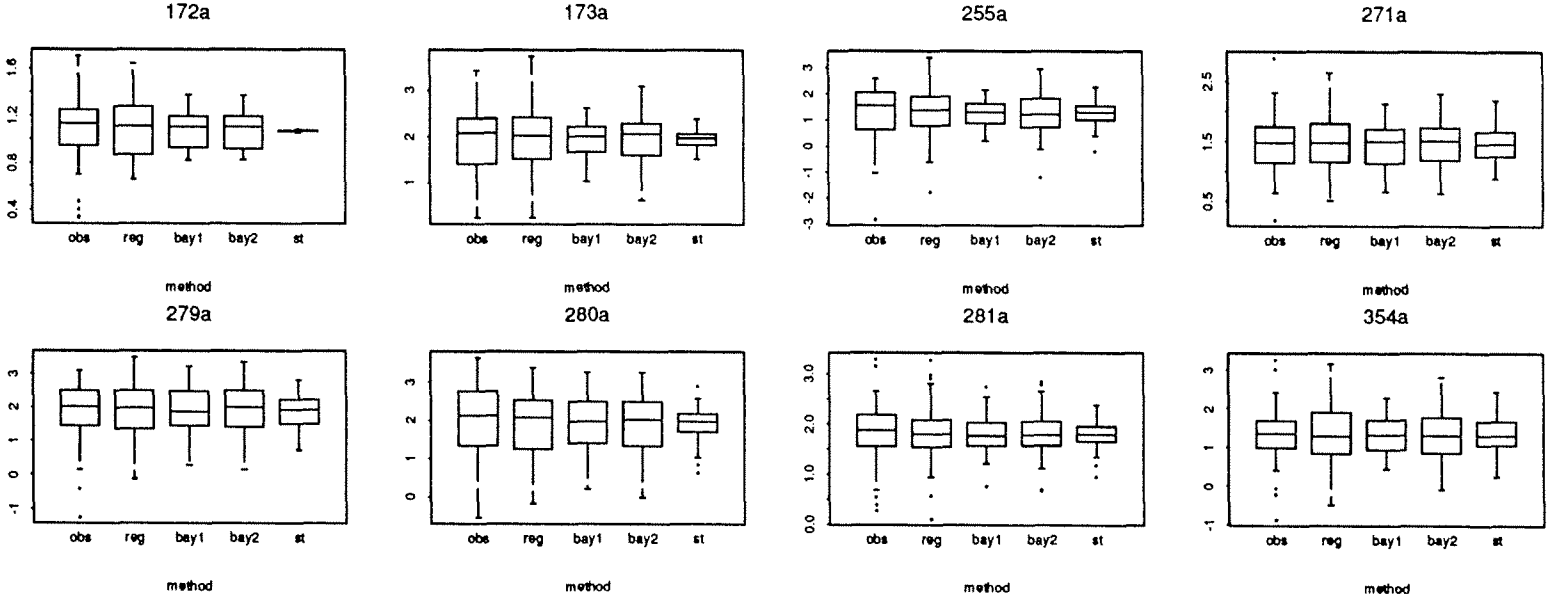
Legend:
obs=observed, reg=ordinary regression, bay1=regression using a Bayesian alternative (1) approach, bay2 = Bayesian alternative (2) approach, st= Stone's procedure

Figure A1.1(b): Boxplots of Observed and Predicted Values for Cluster 2 of Hydrogen



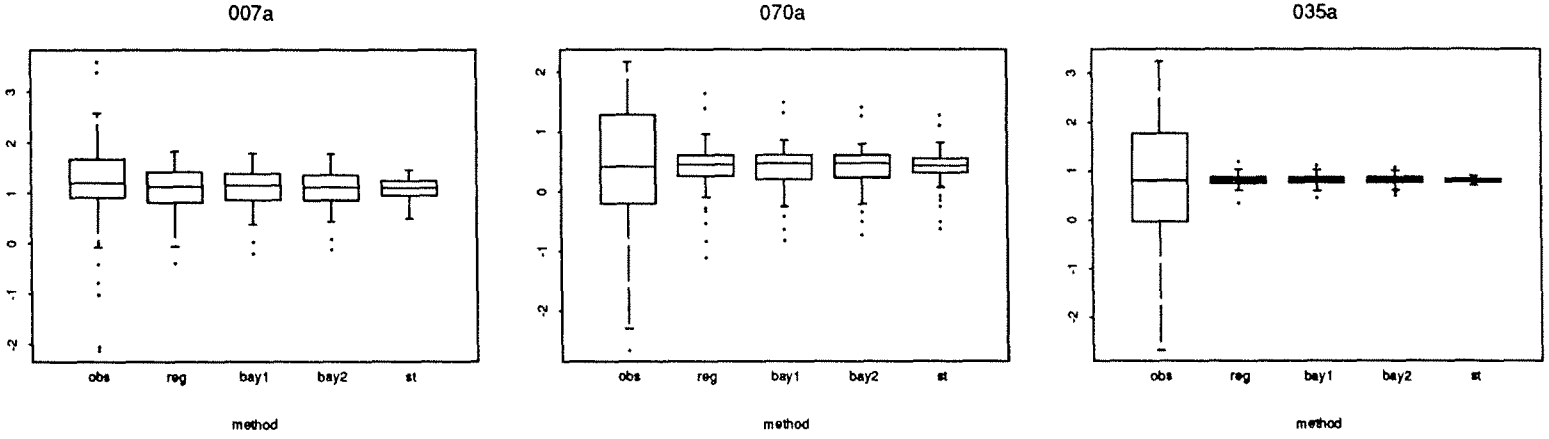
Legend:
obs=observed, reg=ordinary regression, bay1=regression using a Bayesian alternative (1) approach, bay2 = Bayesian alternative (2) approach, st= Stone's procedure

Figure A1.1(b): Continued



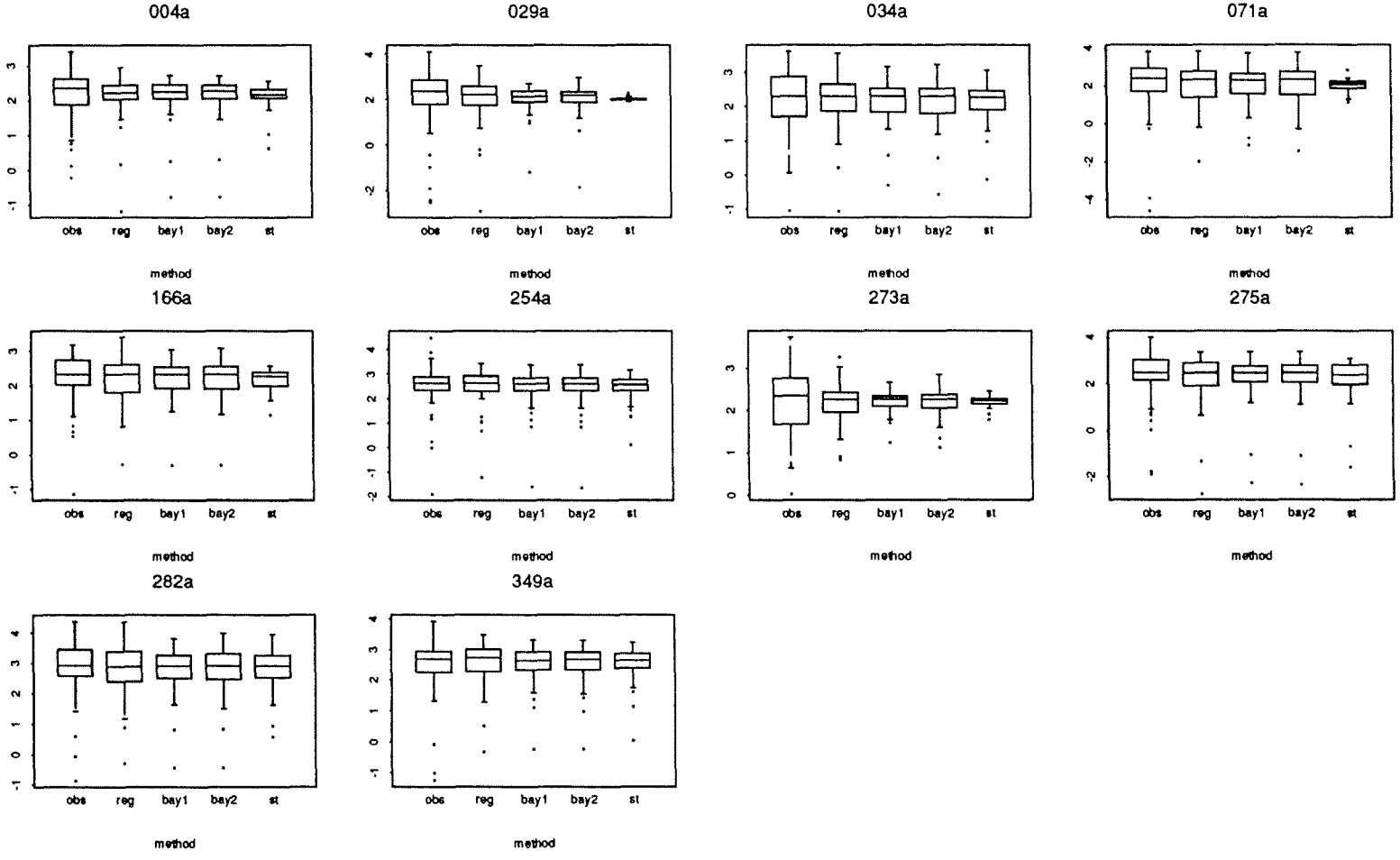
Legend:
obs=observed, reg=ordinary regression, bay1=regression using a Bayesian alternative (1) approach, bay2 = Bayesian alternative (2) approach, st= Stone's procedure

Figure A1.1(c): Boxplots of Observed and Predicted Values for Cluster 3 of Hydrogen



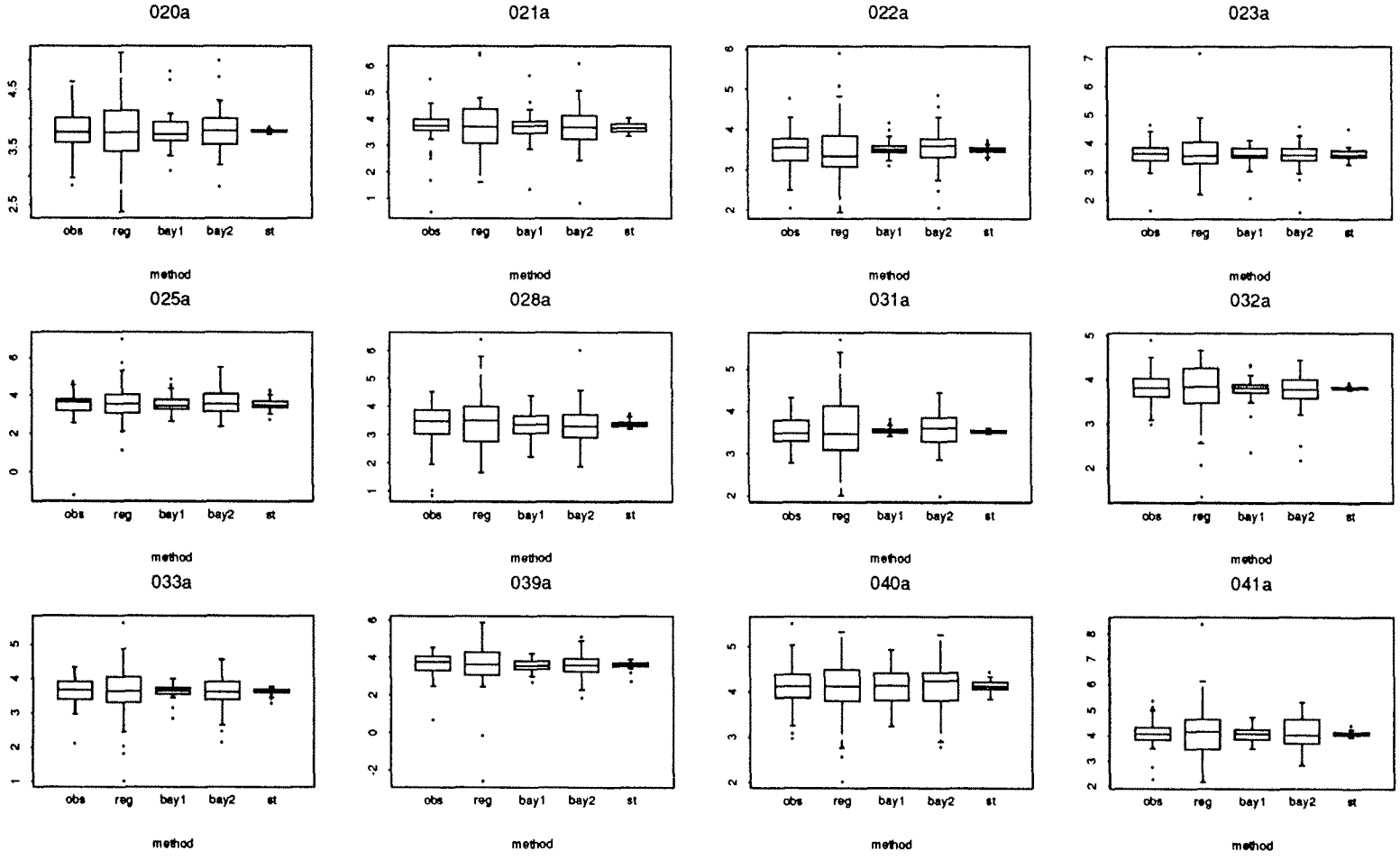
Legend:
obs=observed, reg=ordinary regression, bay1=regression using a Bayesian alternative (1) approach, bay2 = Bayesian alternative (2) approach, st= Stone's procedure

Figure A1.1(d): Boxplots of Observed and Predicted Values for Cluster 4 of Hydrogen



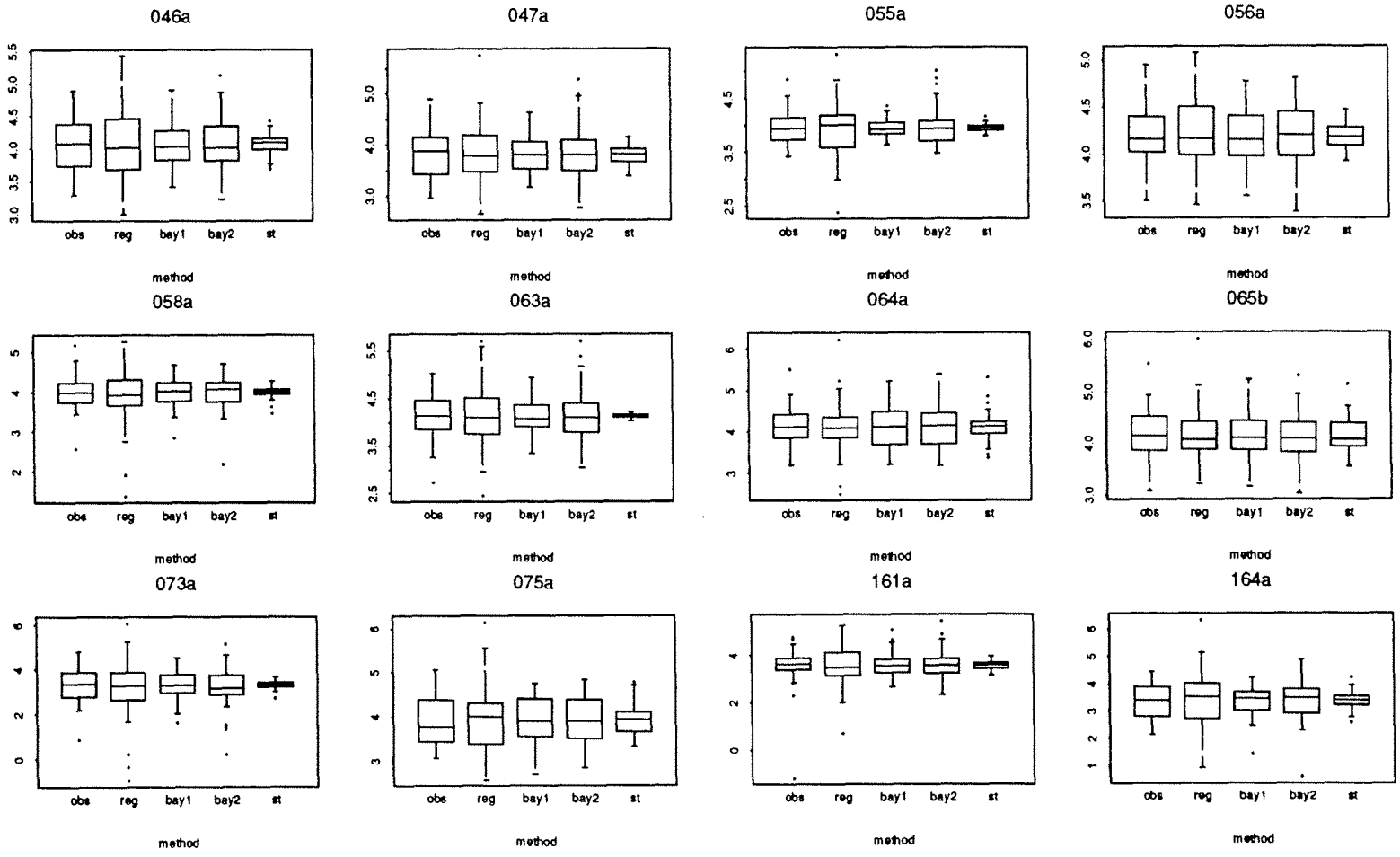
Legend:
obs=observed, reg=ordinary regression, bay1=regression using a Bayesian alternative (1) approach, bay2 = Bayesian alternative (2) approach, st= Stone's procedure

Figure A1.1(e): Boxplots of Observed and Predicted Values for Cluster 5 of Hydrogen



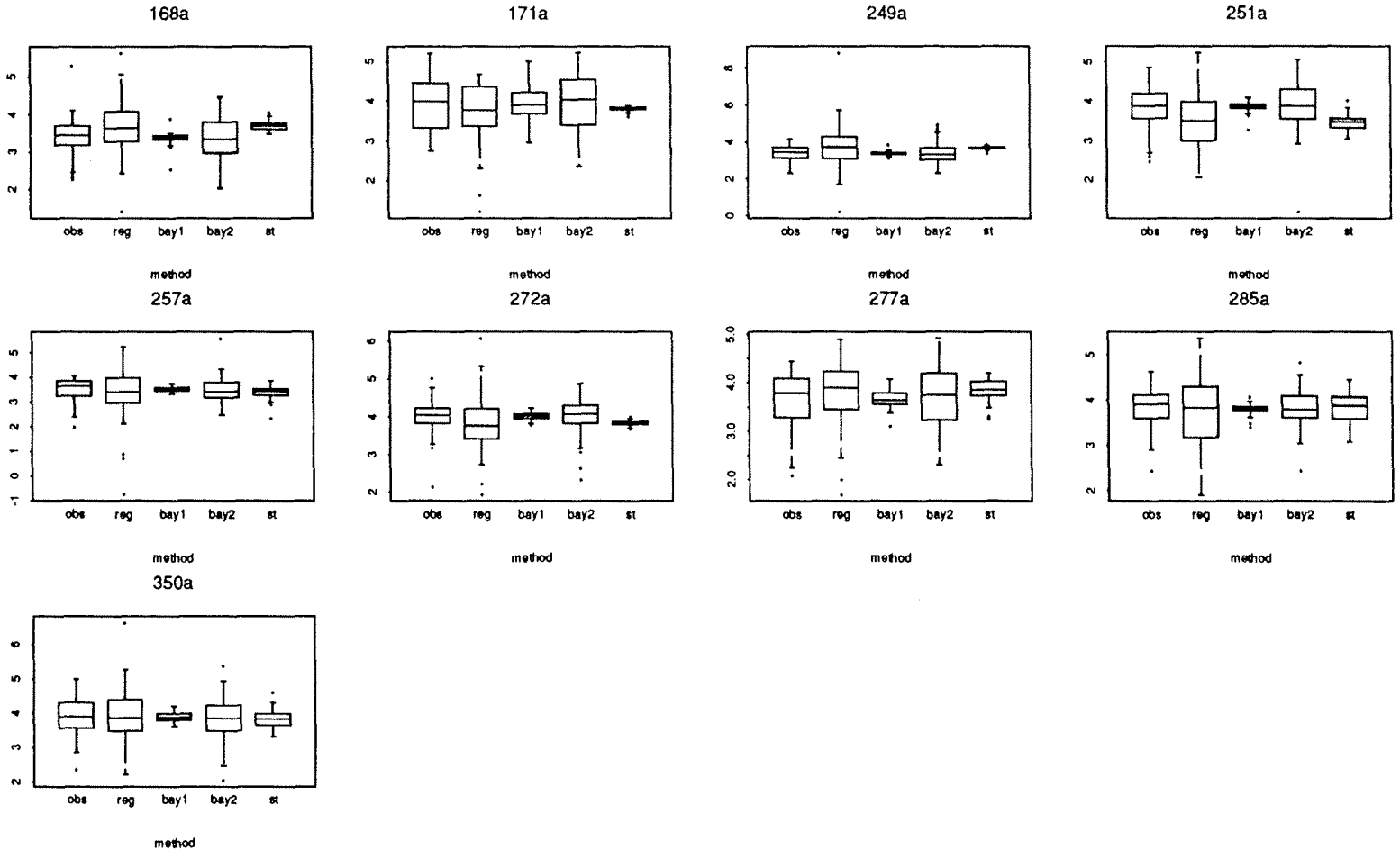
Legend:
obs=observed, reg=ordinary regression, bay1=regression using a Bayesian alternative (1) approach, bay2 = Bayesian alternative (2) approach, st= Stone's procedure

Figure A1.1(e): Continued



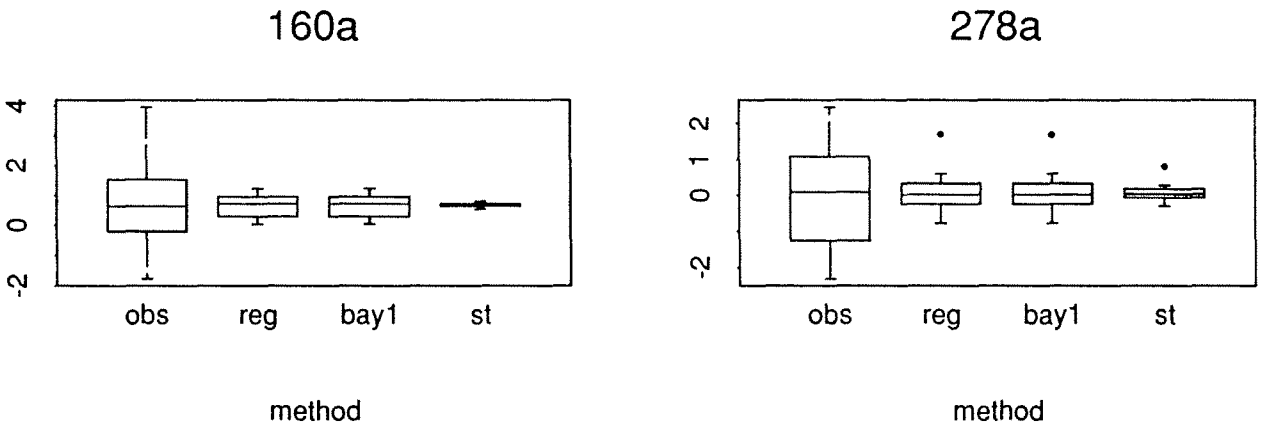
Legend:
obs=observed, reg=ordinary regression, bay1=regression using a Bayesian alternative (1) approach, bay2 = Bayesian alternative (2) approach, st= Stone's procedure

Figure A1.1(e): Continued



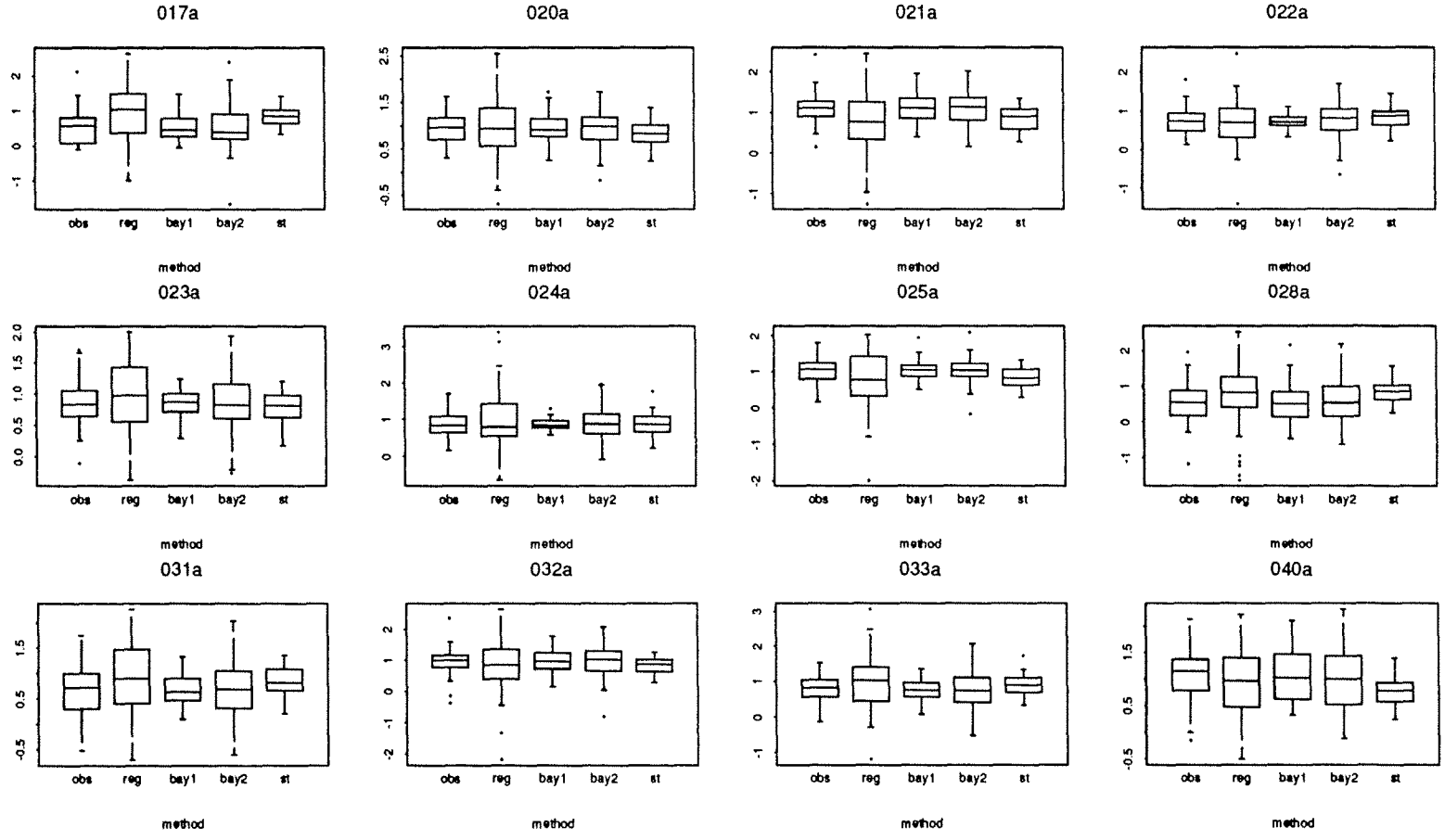
Legend:
obs=observed, reg=ordinary regression, bay1=regression using a Bayesian alternative (1) approach, bay2 = Bayesian alternative (2) approach, st= Stone's procedure

Figure A1.1(f): Boxplots of Observed and Predicted Values for Cluster 6 of Hydrogen



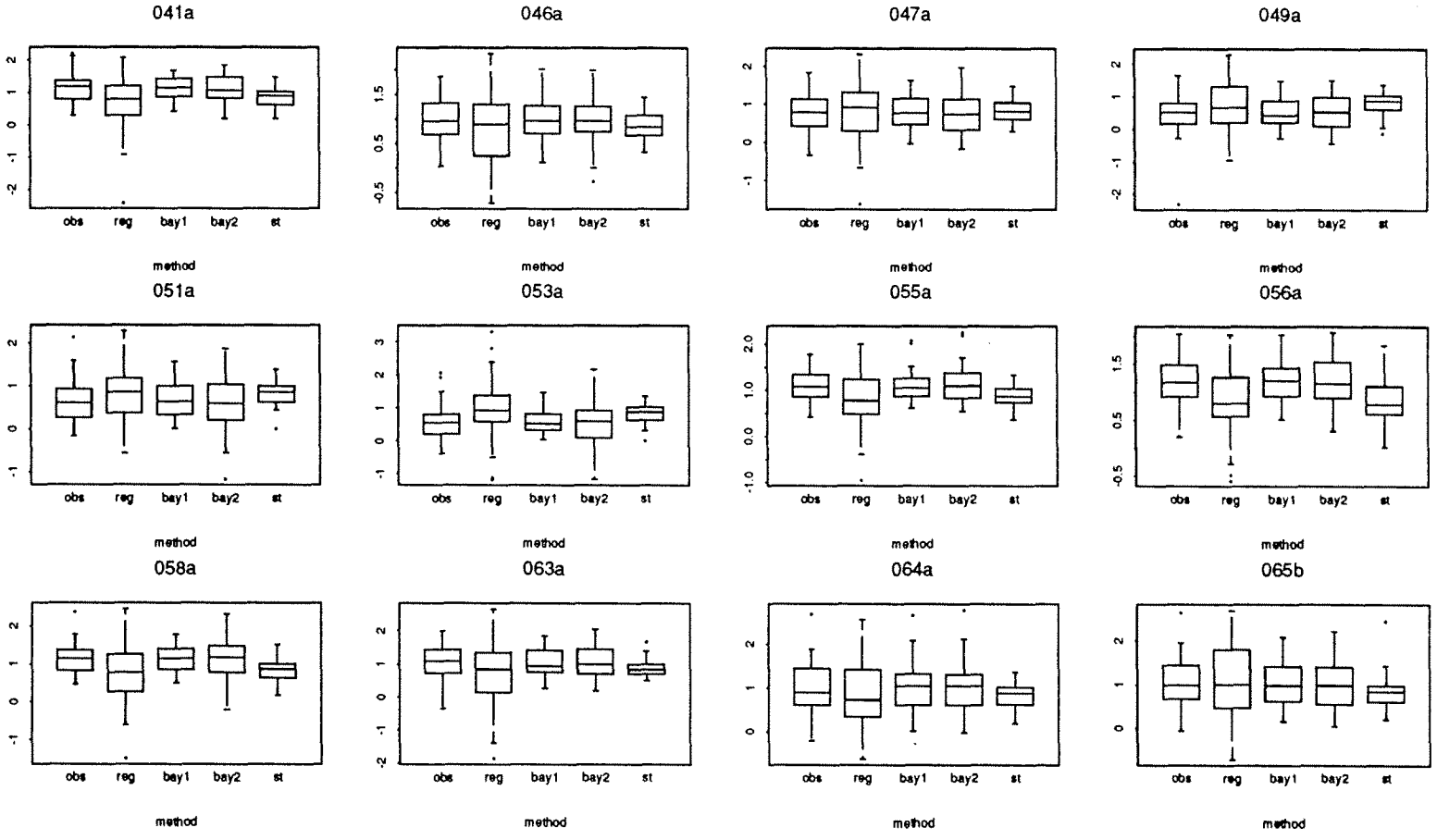
Legend:
obs=observed, reg=ordinary regression, bay1=regression using a Bayesian alternative (1) approach, bay2 = Bayesian alternative (2) approach, st= Stone's procedure

Figure A1.2(a): Boxplots of Observed and Predicted Values for Cluster 1 of Sulfate



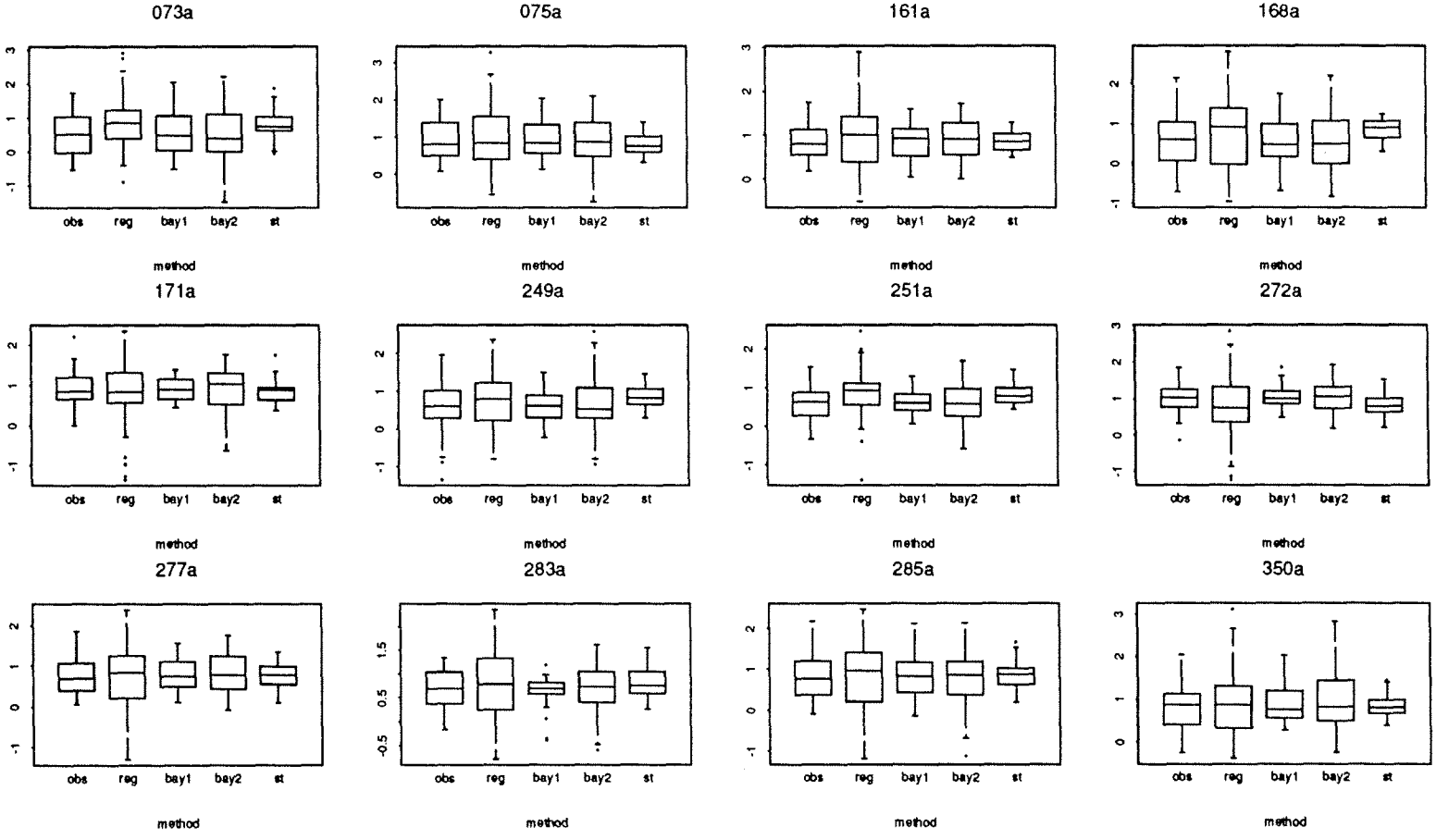
Legend:
obs=observed, reg=ordinary regression, bay1=regression using a Bayesian alternative (1) approach, bay2 = Bayesian alternative (2) approach, st= Stone's procedure

Figure A1.2(a): Continued



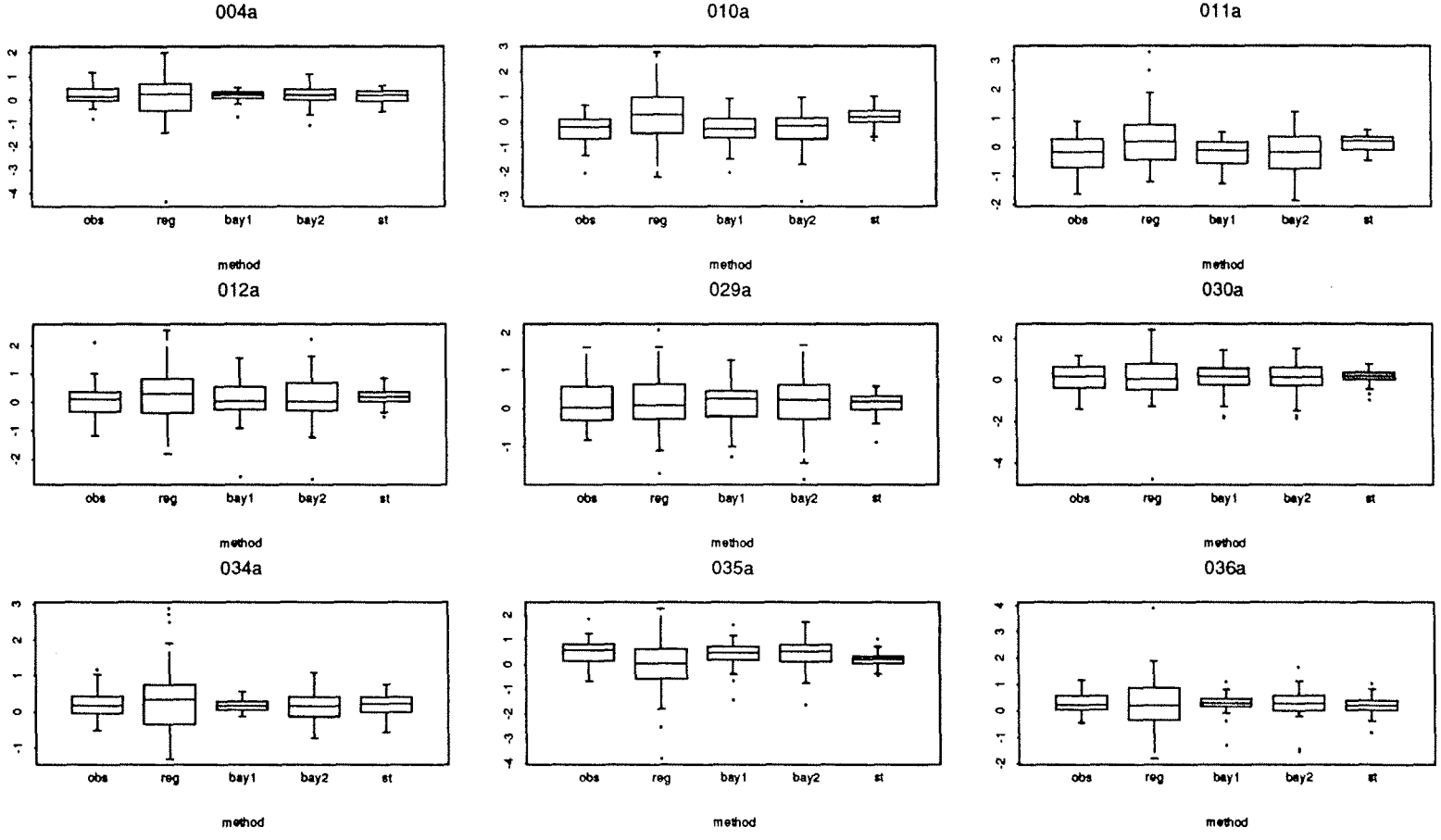
Legend:
obs=observed, reg=ordinary regression, bay1=regression using a Bayesian alternative (1) approach, bay2 = Bayesian alternative (2) approach, st= Stone's procedure

Figure A1.2(a): Continued



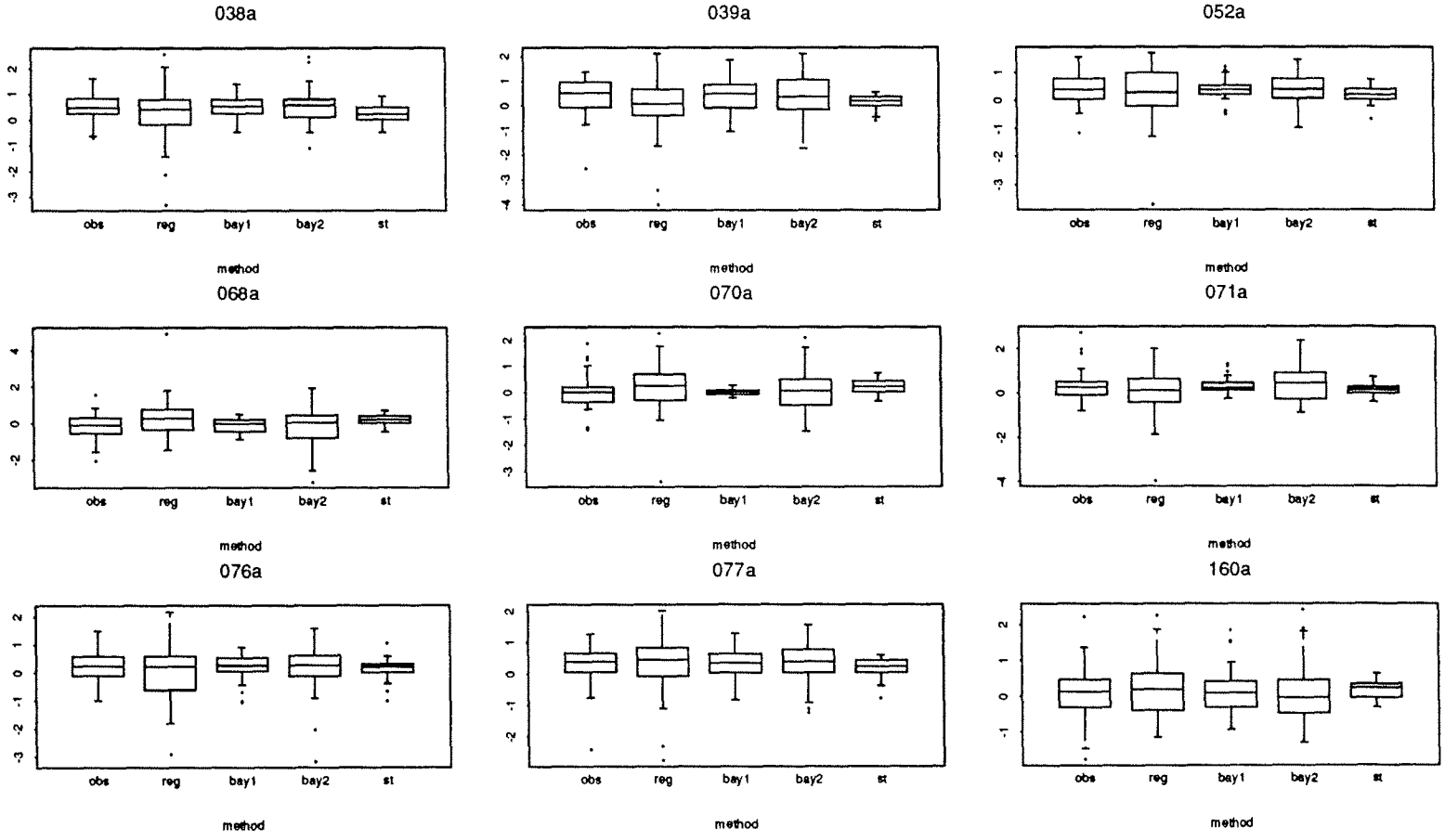
Legend:
obs=observed, reg=ordinary regression, bay1=regression using a Bayesian alternative (1) approach, bay2 = Bayesian alternative (2) approach, st= Stone's procedure

Figure A1.2(b): Boxplots of Observed and Predicted Values for Cluster 2 of Sulfate



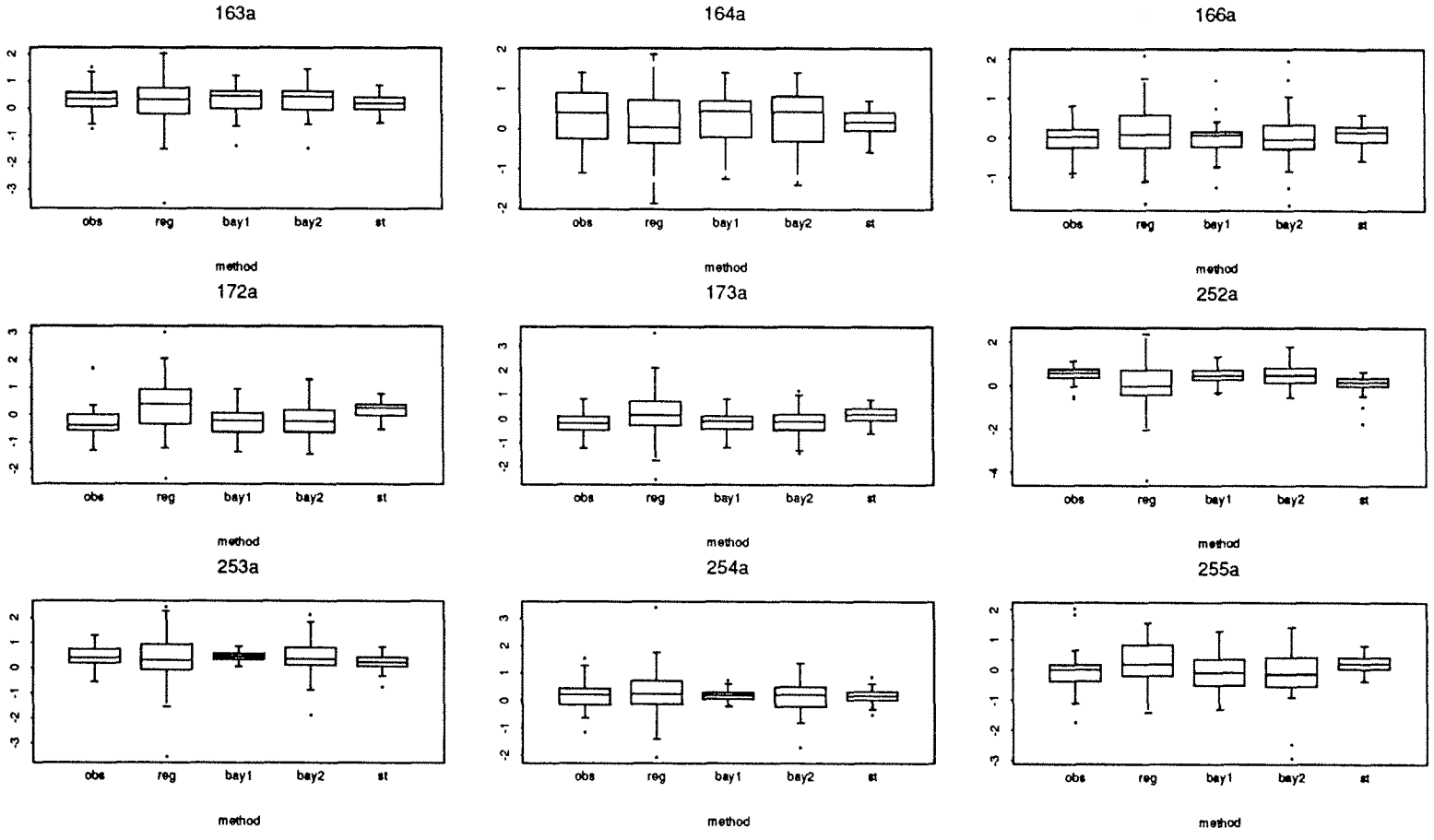
Legend:
obs=observed, reg=ordinary regression, bay1=regression using a Bayesian alternative (1) approach, bay2 = Bayesian alternative (2) approach, st= Stone's procedure

Figure A1.2(b): Continued



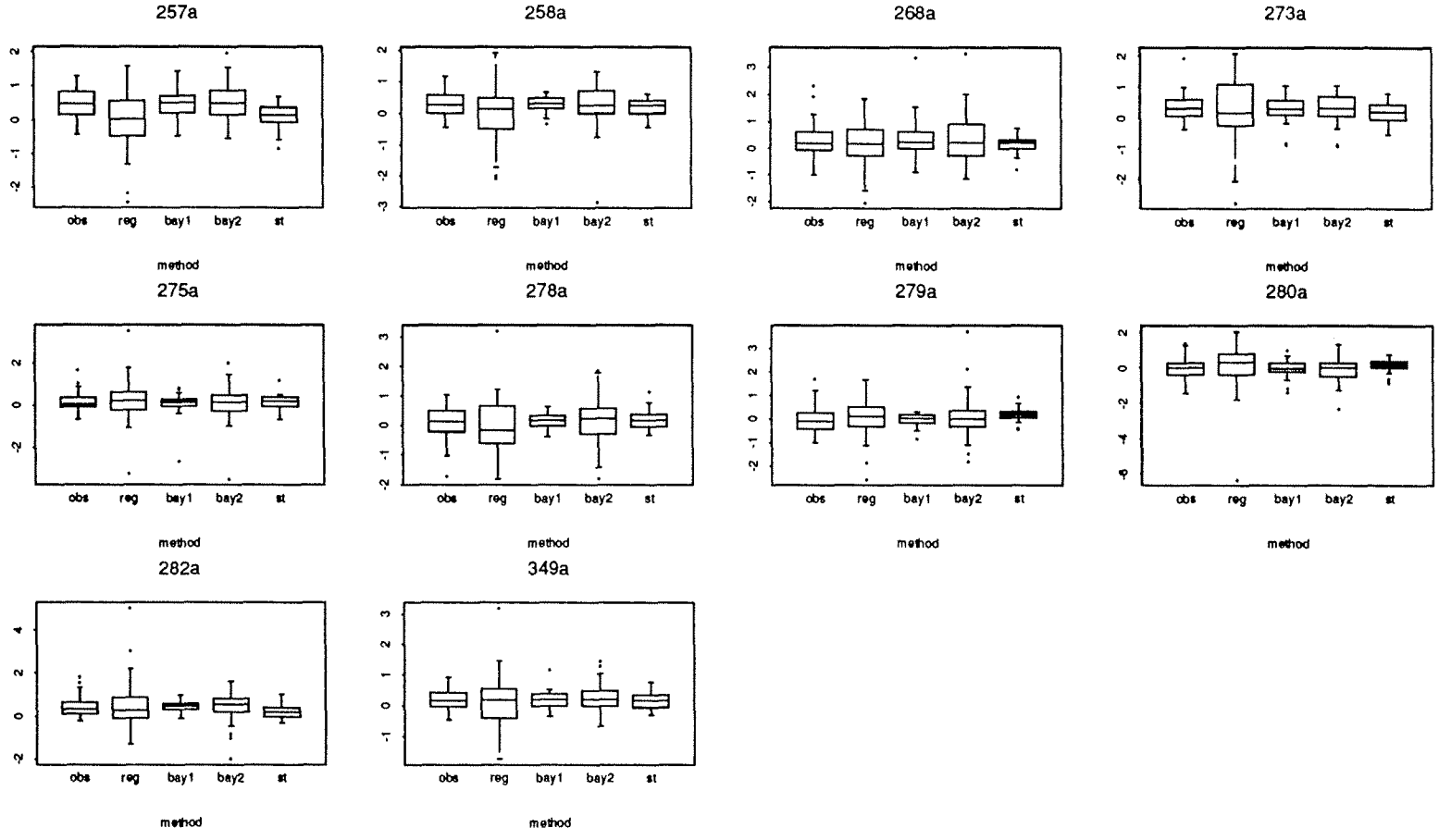
Legend:
obs=observed, reg=ordinary regression, bay1=regression using a Bayesian alternative (1) approach, bay2 = Bayesian alternative (2) approach, st= Stone's procedure

Figure A1.2(b): Continued



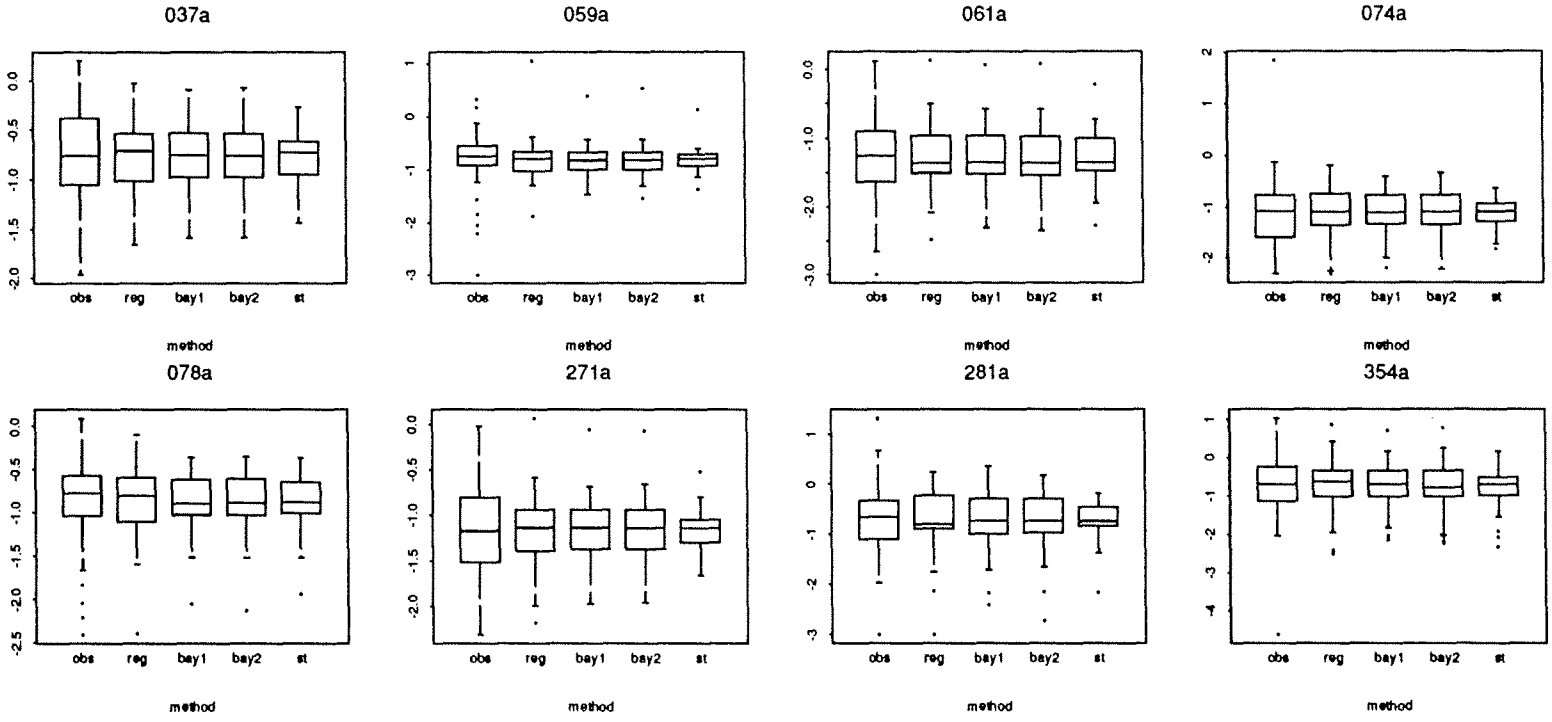
Legend:
obs=observed, reg=ordinary regression, bay1=regression using a Bayesian alternative (1) approach, bay2 = Bayesian alternative (2) approach, st= Stone's procedure

Figure A1.2(b): Continued



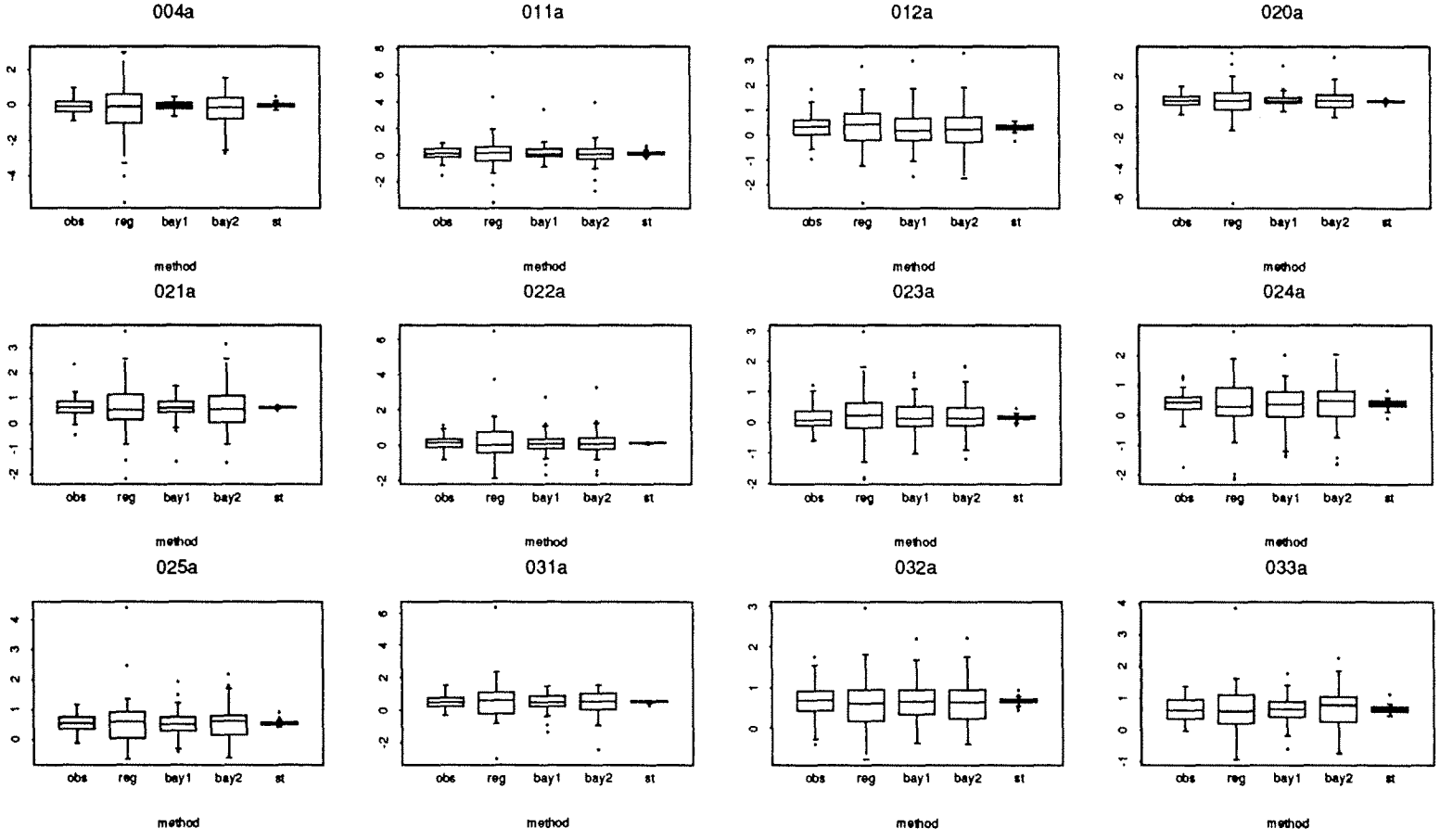
Legend:
obs=observed, reg=ordinary regression, bay1=regression using a Bayesian alternative (1) approach, bay2 = Bayesian alternative (2) approach, st= Stone's procedure

Figure A1.2(c): Boxplots of Observed and Predicted Values for Cluster 3 of Sulfate



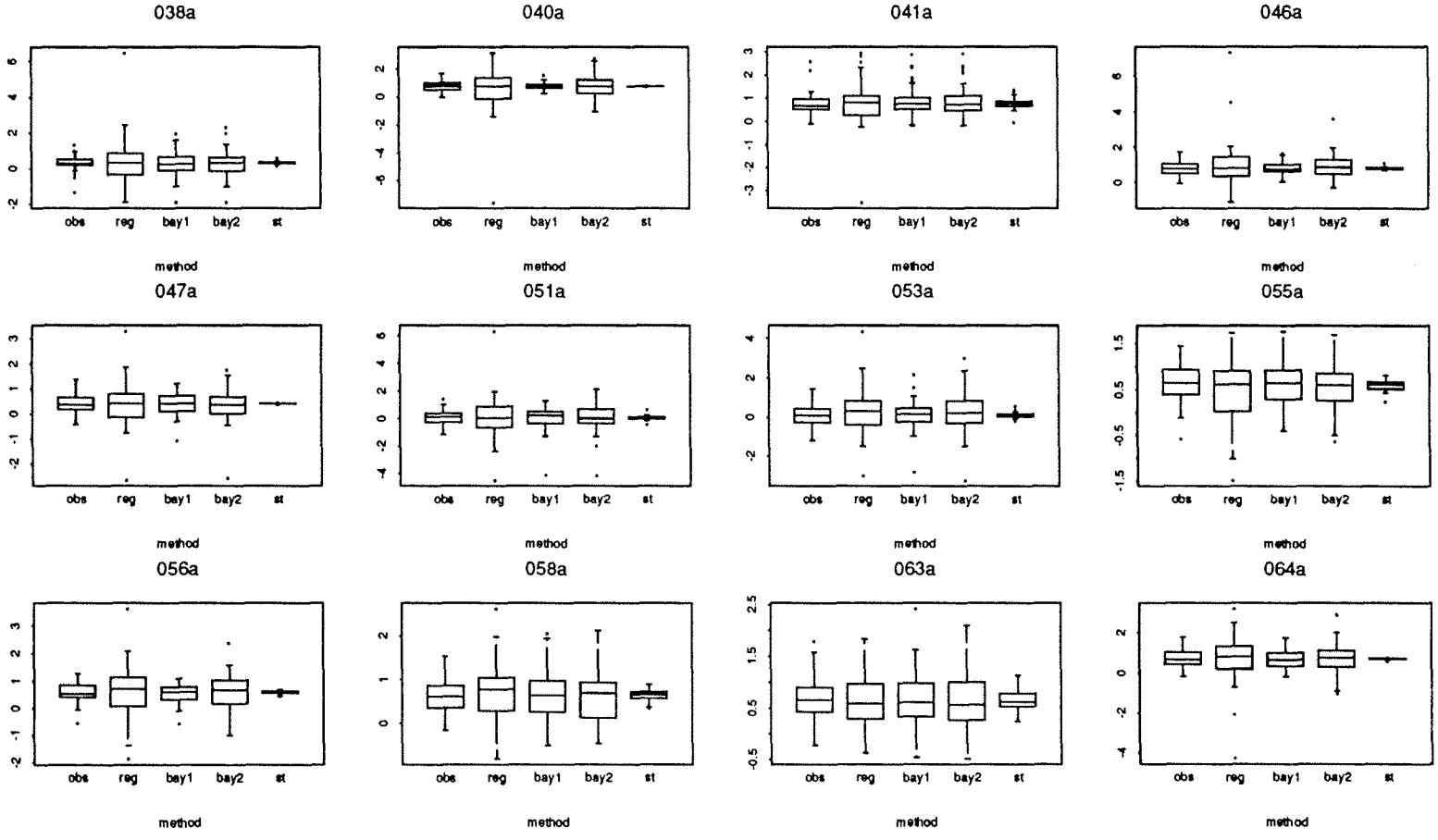
Legend:
obs=observed, reg=ordinary regression, bay1=regression using a Bayesian alternative (1) approach, bay2 = Bayesian alternative (2) approach, st= Stone's procedure

Figure A1.3(a): Boxplots of Observed and Predicted Values for Cluster 1 of Nitrate



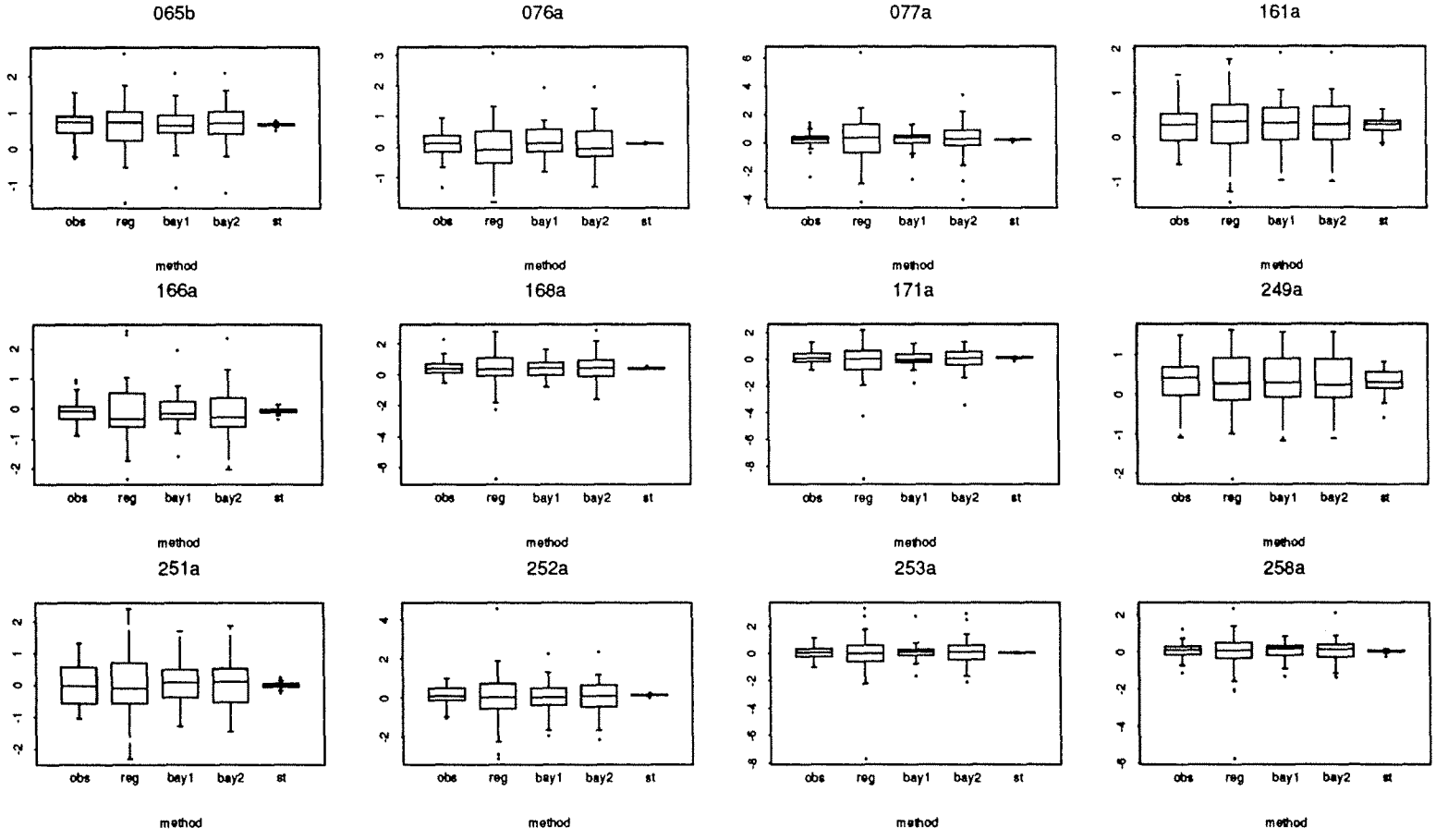
Legend:
obs=observed, reg=ordinary regression, bay1=regression using a Bayesian alternative (1) approach, bay2 = Bayesian alternative (2) approach, st= Stone's procedure

Figure A1.3(a): Continued



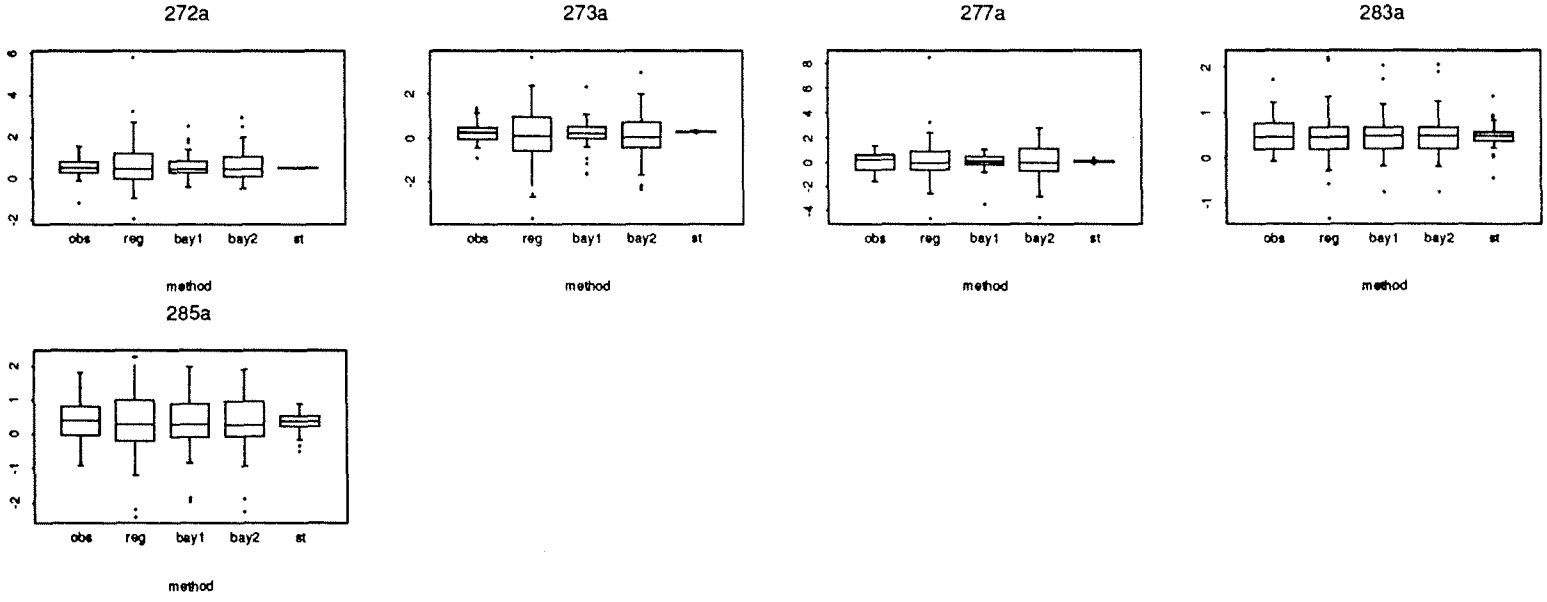
Legend:
obs=observed, reg=ordinary regression, bay1=regression using a Bayesian alternative (1) approach, bay2 = Bayesian alternative (2) approach, st= Stone's procedure

Figure A1.3(a): Continued



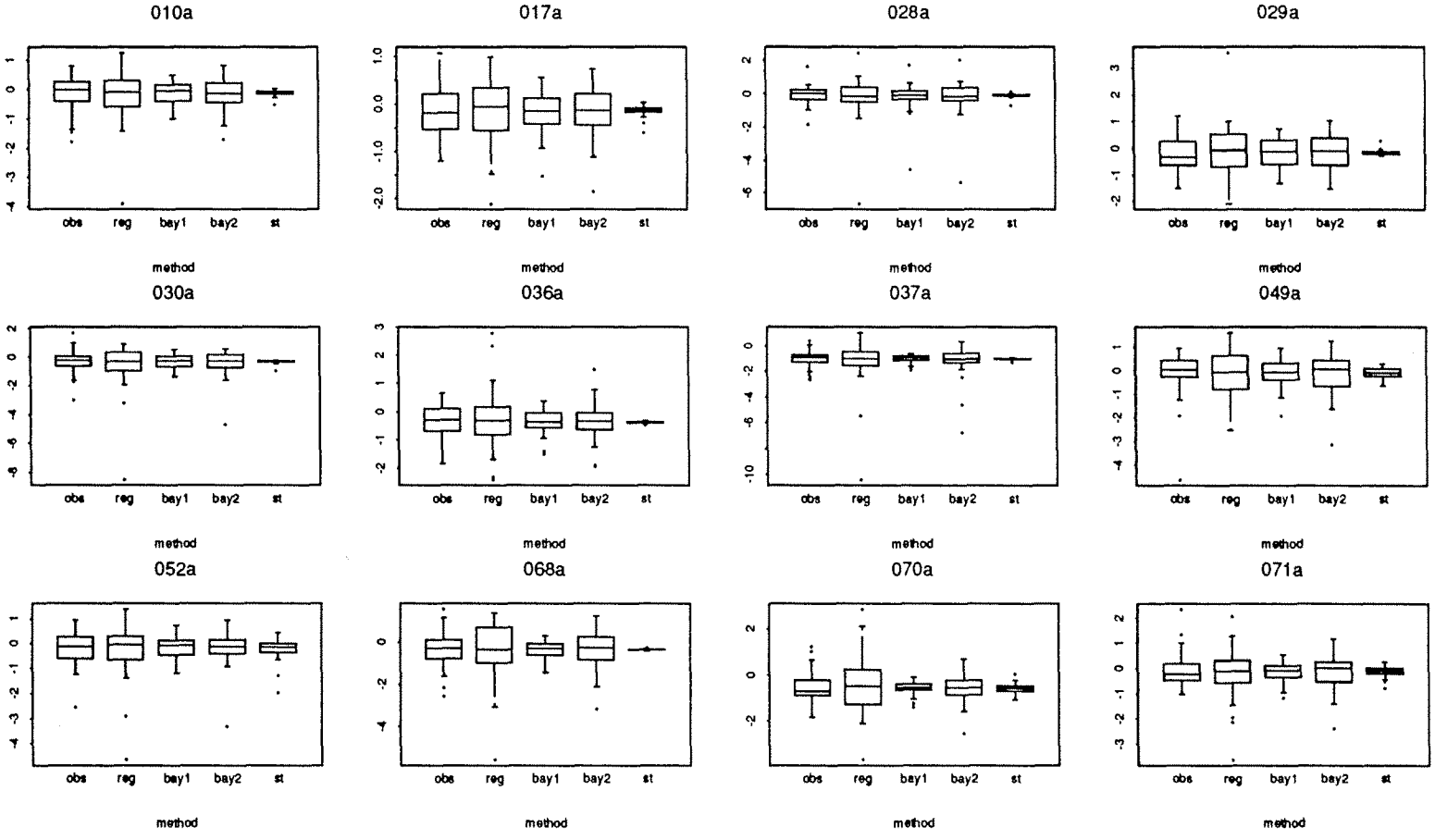
Legend:
obs=observed, reg=ordinary regression, bay1=regression using a Bayesian alternative (1) approach, bay2 = Bayesian alternative (2) approach, st= Stone's procedure

Figure A1.3(a): Continued



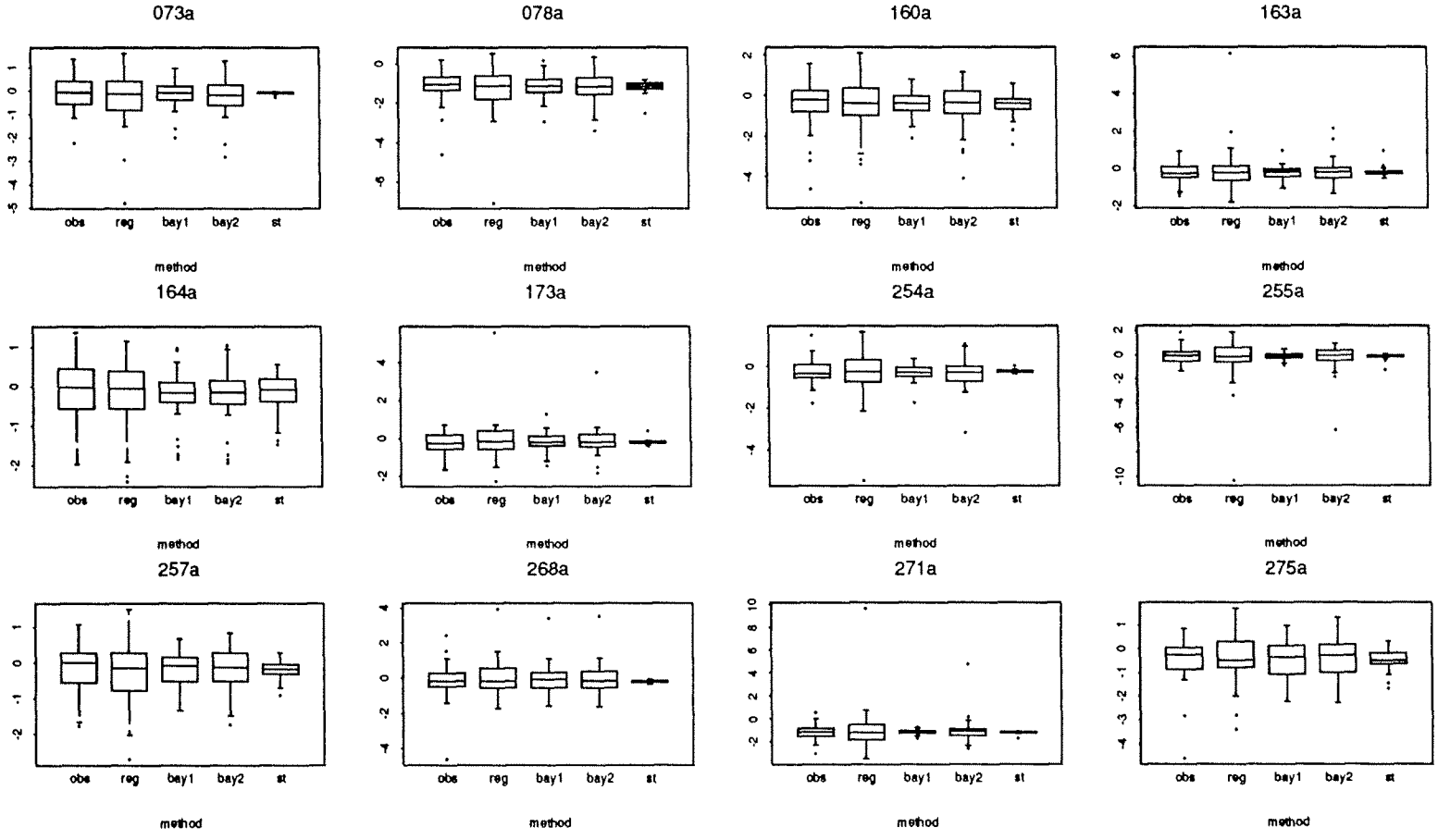
Legend:
obs=observed, reg=ordinary regression, bay1=regression using a Bayesian alternative (1) approach, bay2 = Bayesian alternative (2) approach, st= Stone's procedure

Figure A1.3(b): Boxplots of Observed and Predicted Values for Cluster 2 of Nitrate



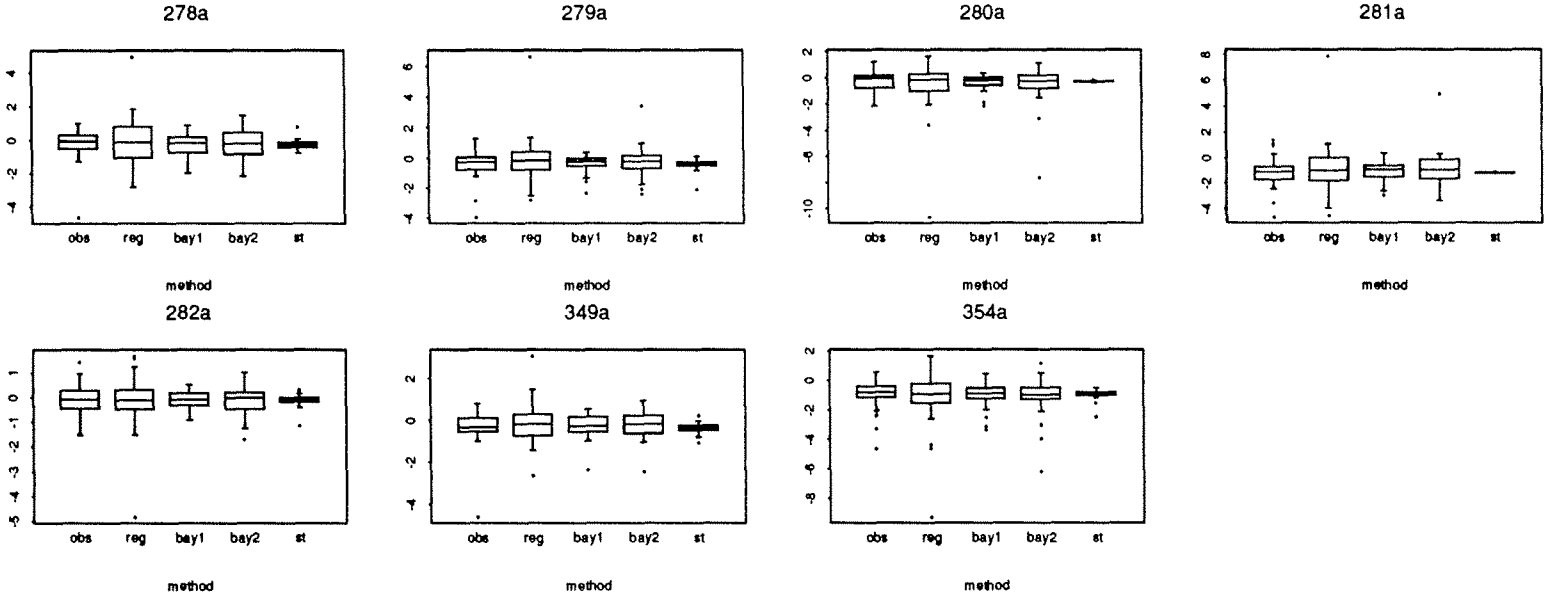
Legend:
obs=observed, reg=ordinary regression, bay1=regression using a Bayesian alternative (1) approach, bay2 = Bayesian alternative (2) approach, st= Stone's procedure

Figure A1.3(b): Continued



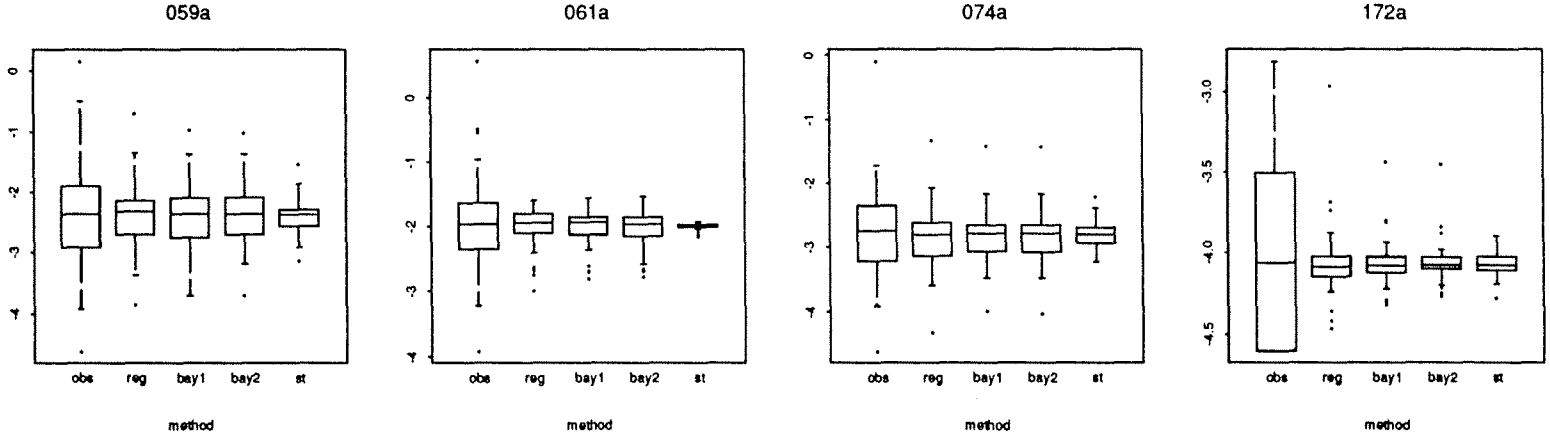
Legend:
obs=observed, reg=ordinary regression, bay1=regression using a Bayesian alternative (1) approach, bay2 = Bayesian alternative (2) approach, st= Stone's procedure

Figure A1.3(b): Continued



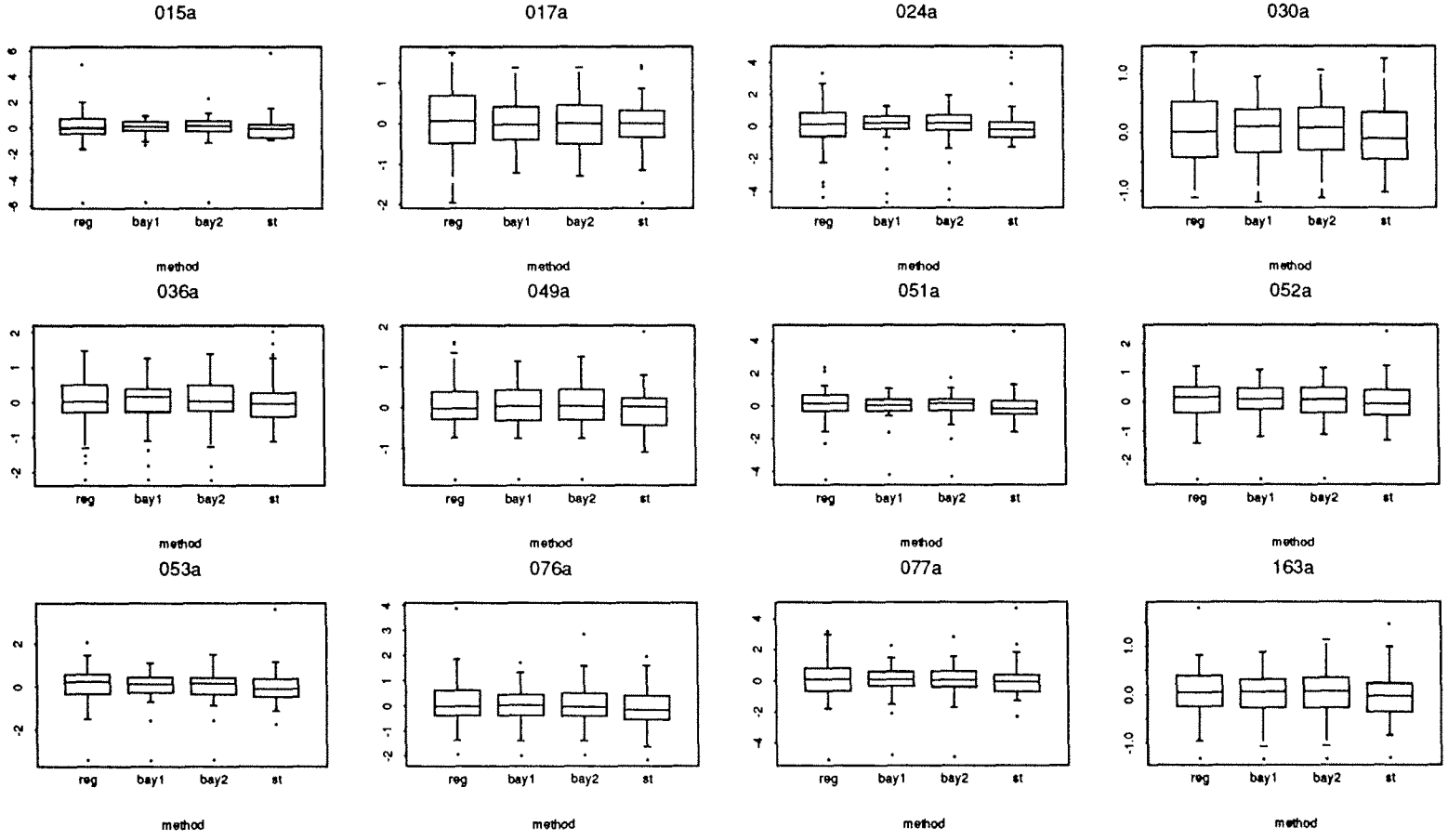
Legend:
obs=observed, reg=ordinary regression, bay1=regression using a Bayesian alternative (1) approach, bay2 = Bayesian alternative (2) approach, st= Stone's procedure

Figure A1.3(c): Boxplots of Observed and Predicted Values for Cluster 3 of Nitrate



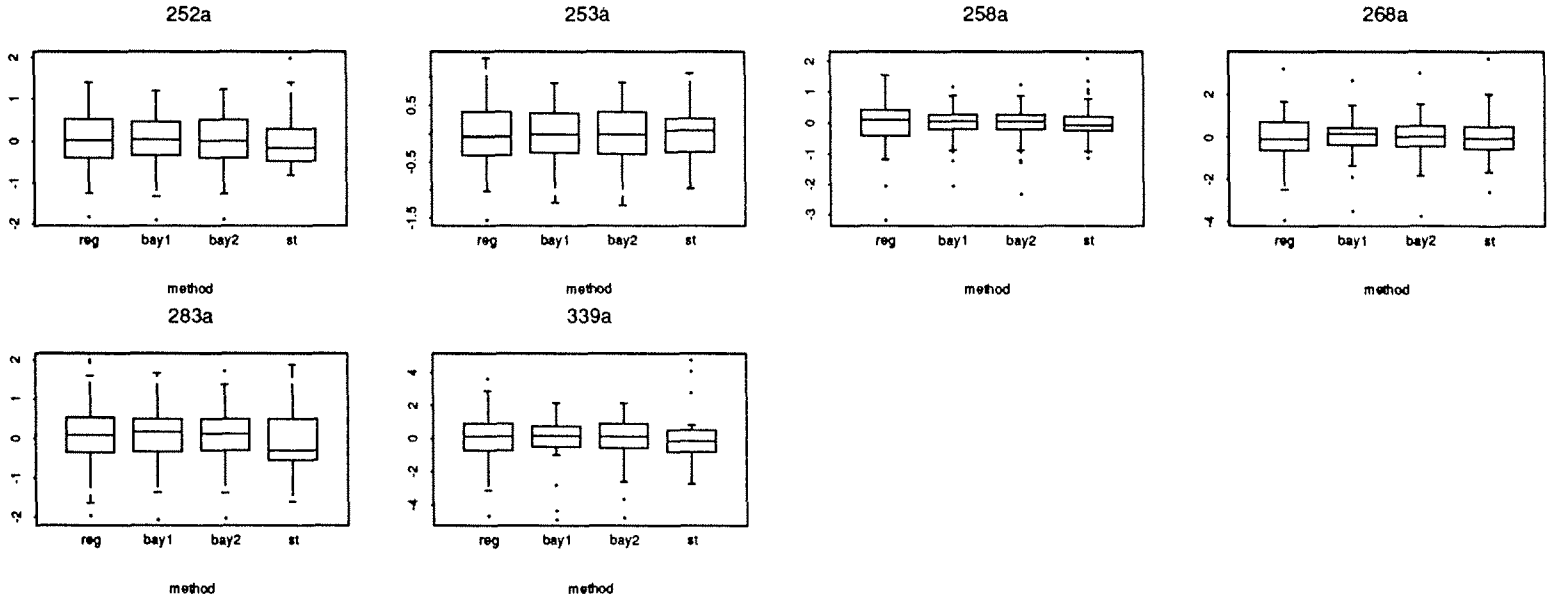
Legend:
obs=observed, reg=ordinary regression, bay1=regression using a Bayesian alternative (1) approach, bay2 = Bayesian alternative (2) approach, st= Stone's procedure

Figure A2.1(a): Boxplots of Prediction Errors for Cluster 1 of Hydrogen



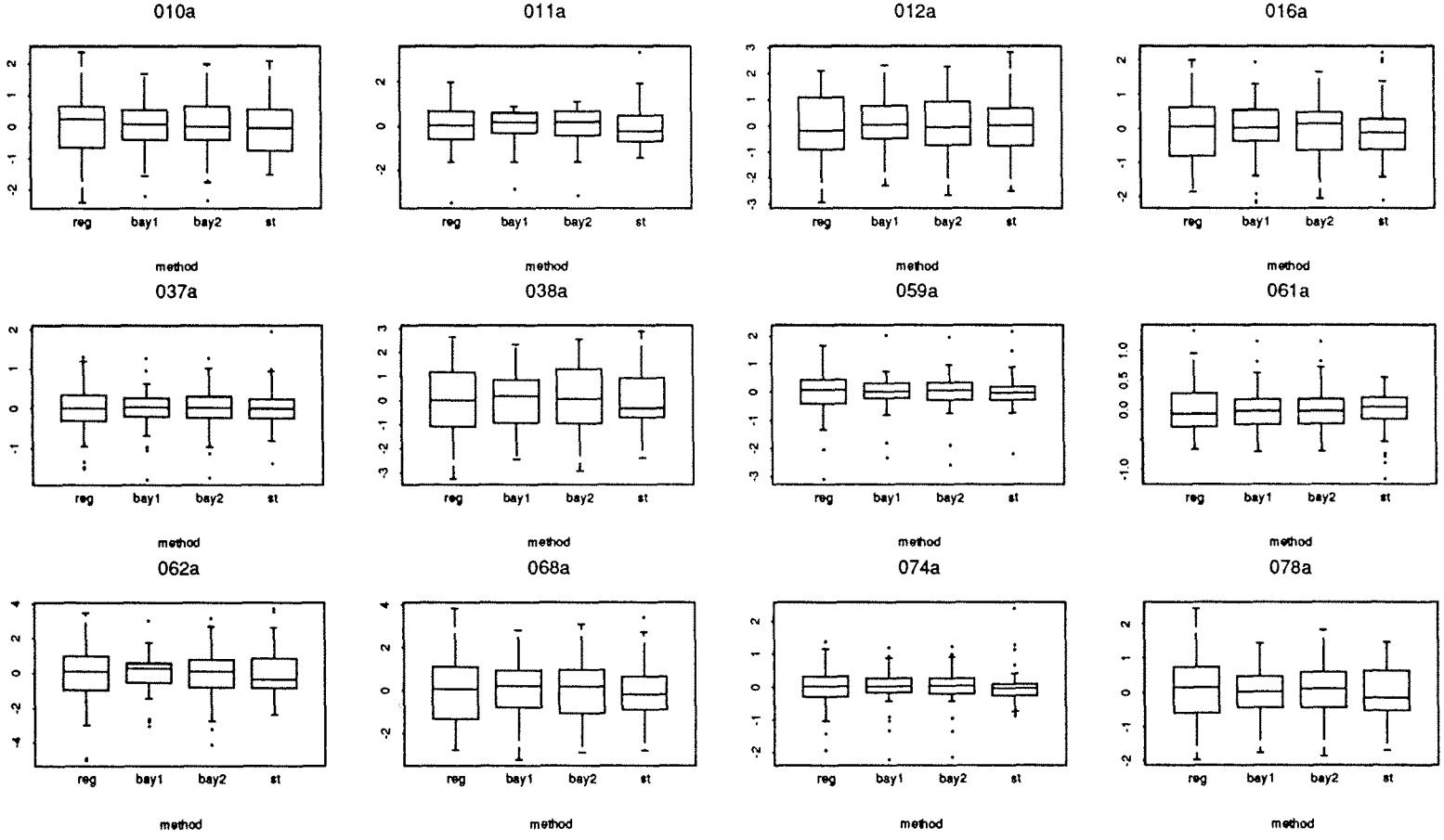
Legend:
reg=ordinary regression, bay1=regression using a Bayesian alternative (1) approach,
bay2= regression using a Bayesian alternative (2) approach, st= Stone's procedure

Figure A2.1(a): Continued



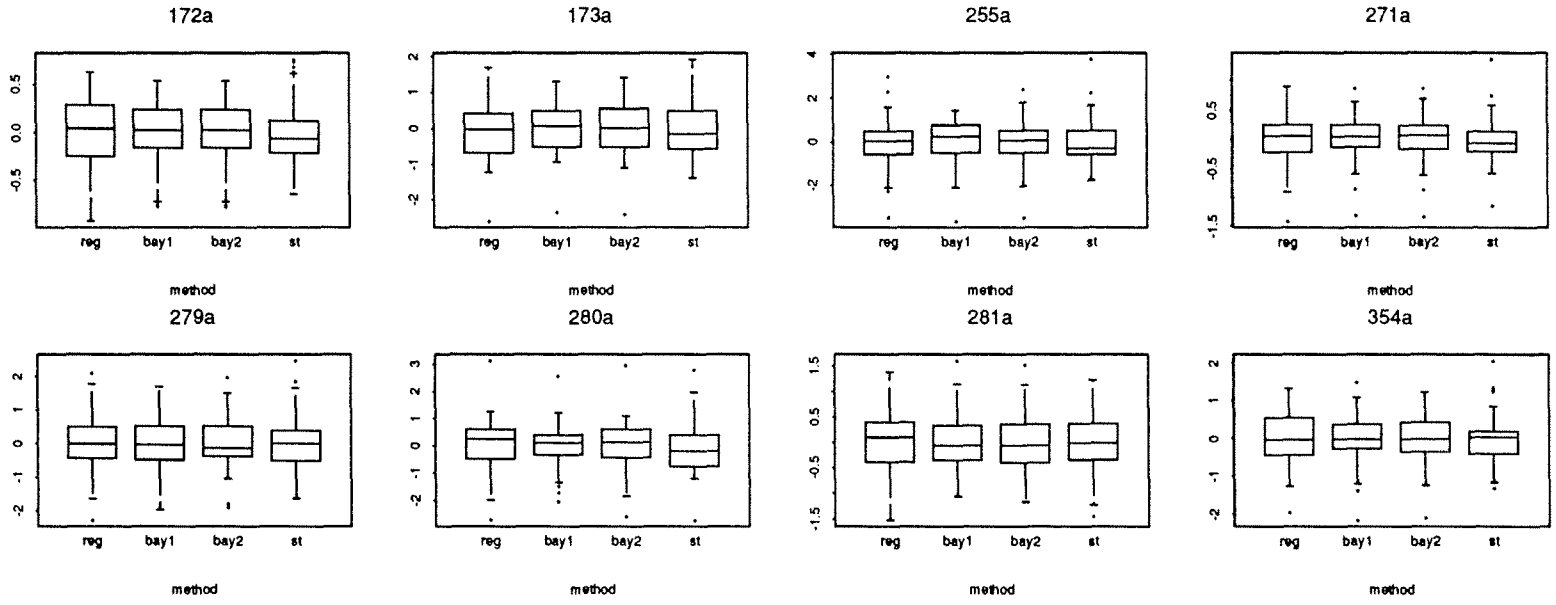
Legend:
reg=ordinary regression, bay1=regression using a Bayesian alternative (1) approach,
bay2= regression using a Bayesian alternative (2) approach, st= Stone's procedure

Figure A2.1(b): Boxplots of Prediction Errors for Cluster 2 of Hydrogen



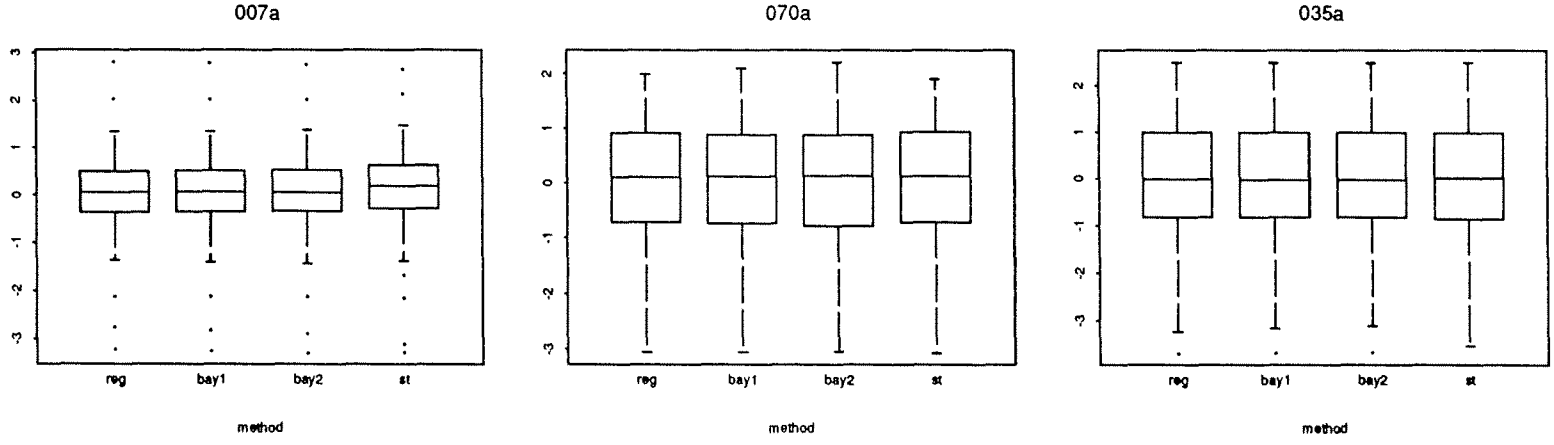
Legend:
reg=ordinary regression, bay1=regression using a Bayesian alternative (1) approach,
bay2= regression using a Bayesian alternative (2) approach, st= Stone's procedure

Figure A2.1(b): Continued



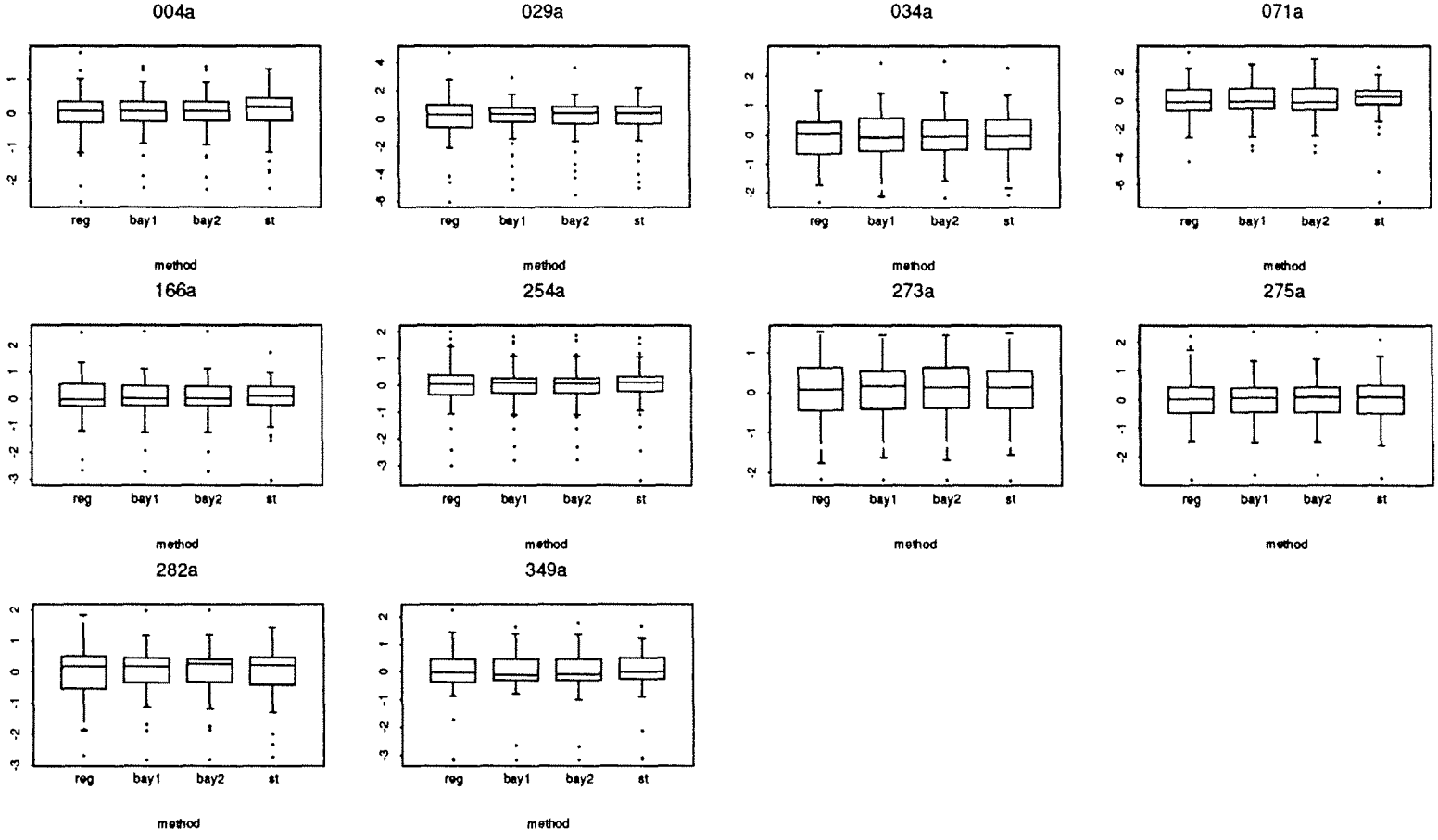
Legend:
reg=ordinary regression, bay1=regression using a Bayesian alternative (1) approach,
bay2= regression using a Bayesian alternative (2) approach, st= Stone's procedure

Figure A2.1(c): Boxplots of Prediction Errors for Cluster 3 of Hydrogen



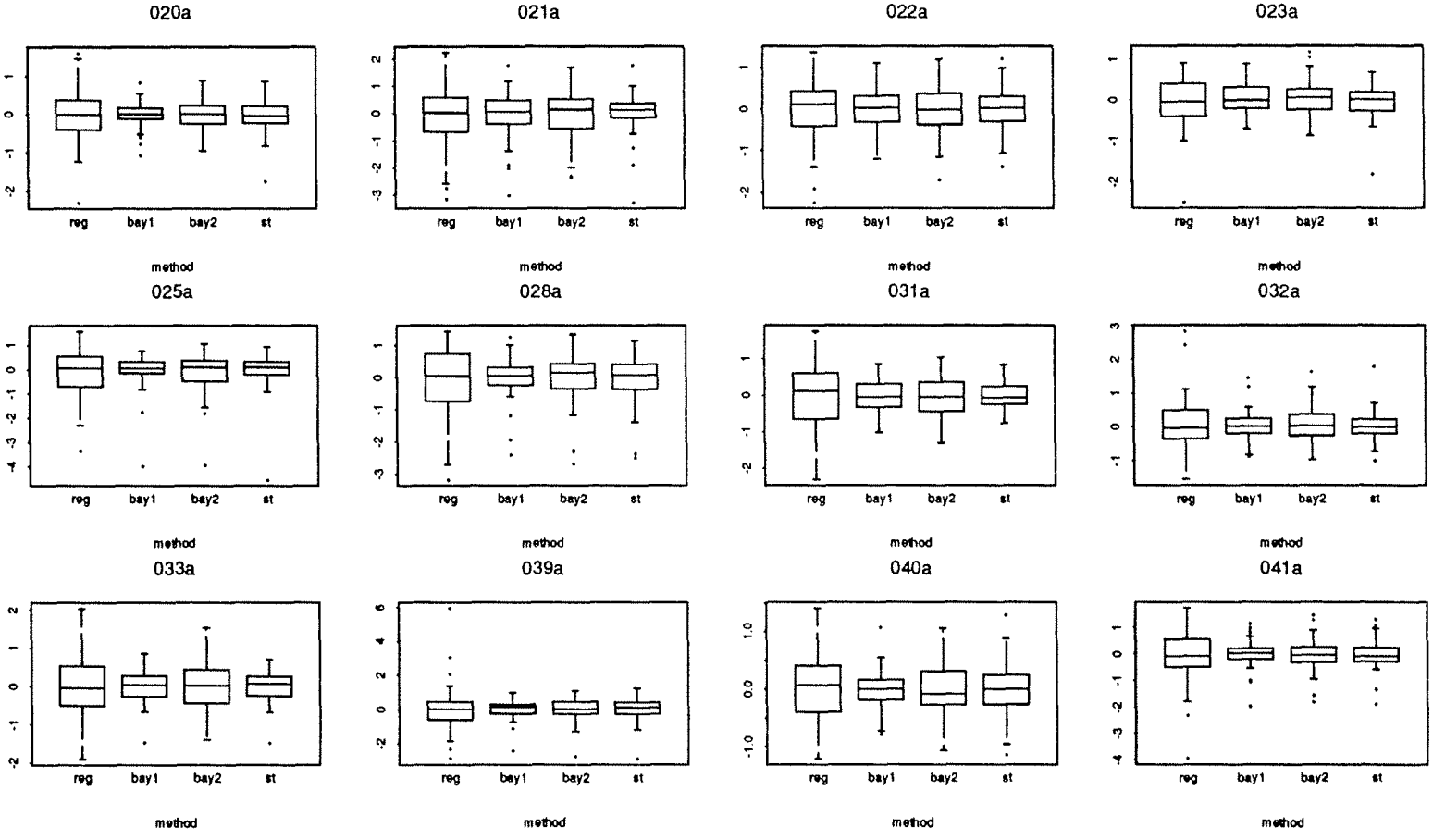
Legend:
reg=ordinary regression, bay1=regression using a Bayesian alternative (1) approach,
bay2= regression using a Bayesian alternative (2) approach, st= Stone's procedure

Figure A2.1(d): Boxplots of Prediction Errors for Cluster 4 of Hydrogen



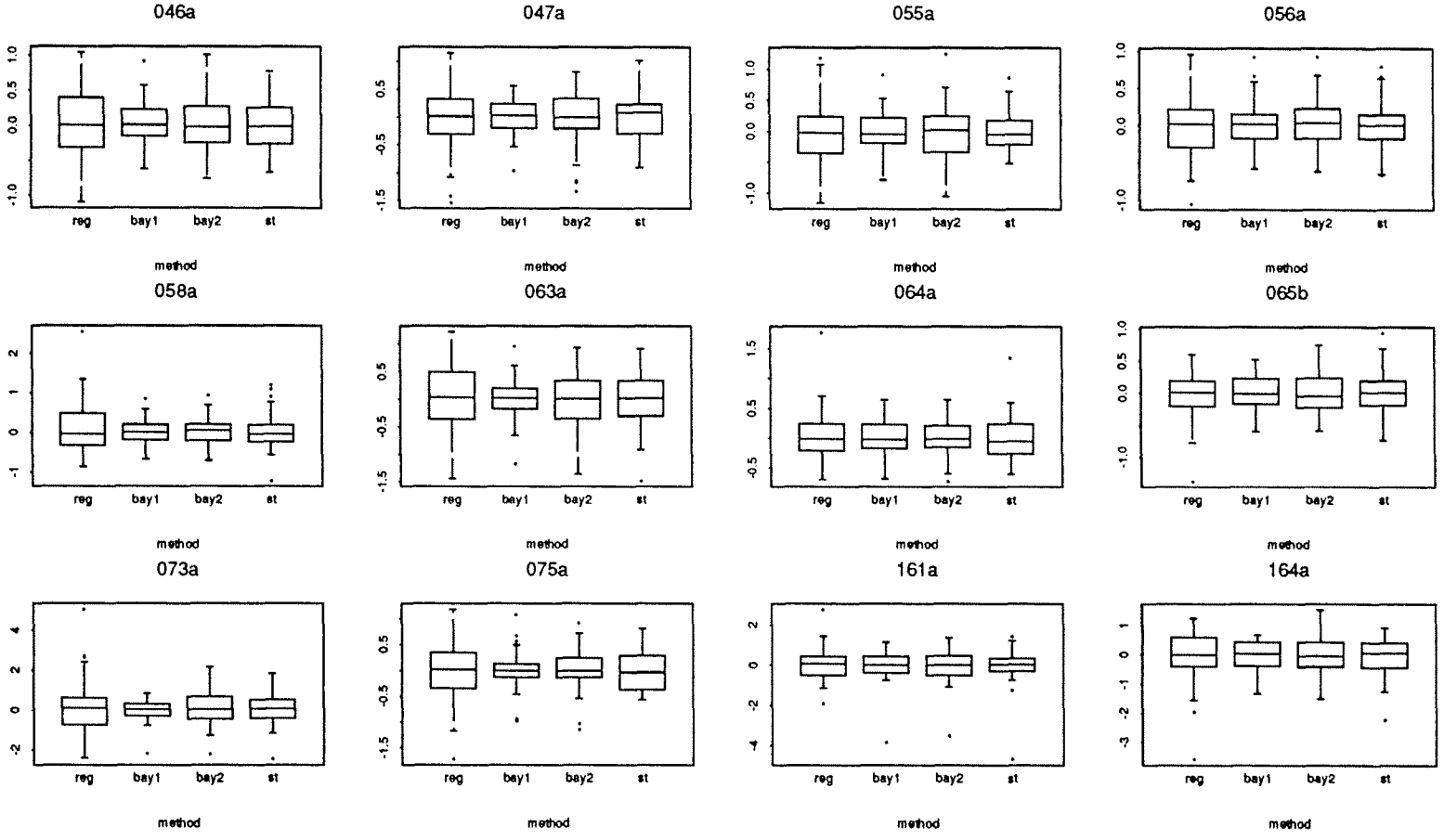
Legend:
reg=ordinary regression, bay1=regression using a Bayesian alternative (1) approach,
bay2= regression using a Bayesian alternative (2) approach, st= Stone's procedure

Figure A2.1(e): Boxplots of Prediction Errors for Cluster 5 of Hydrogen



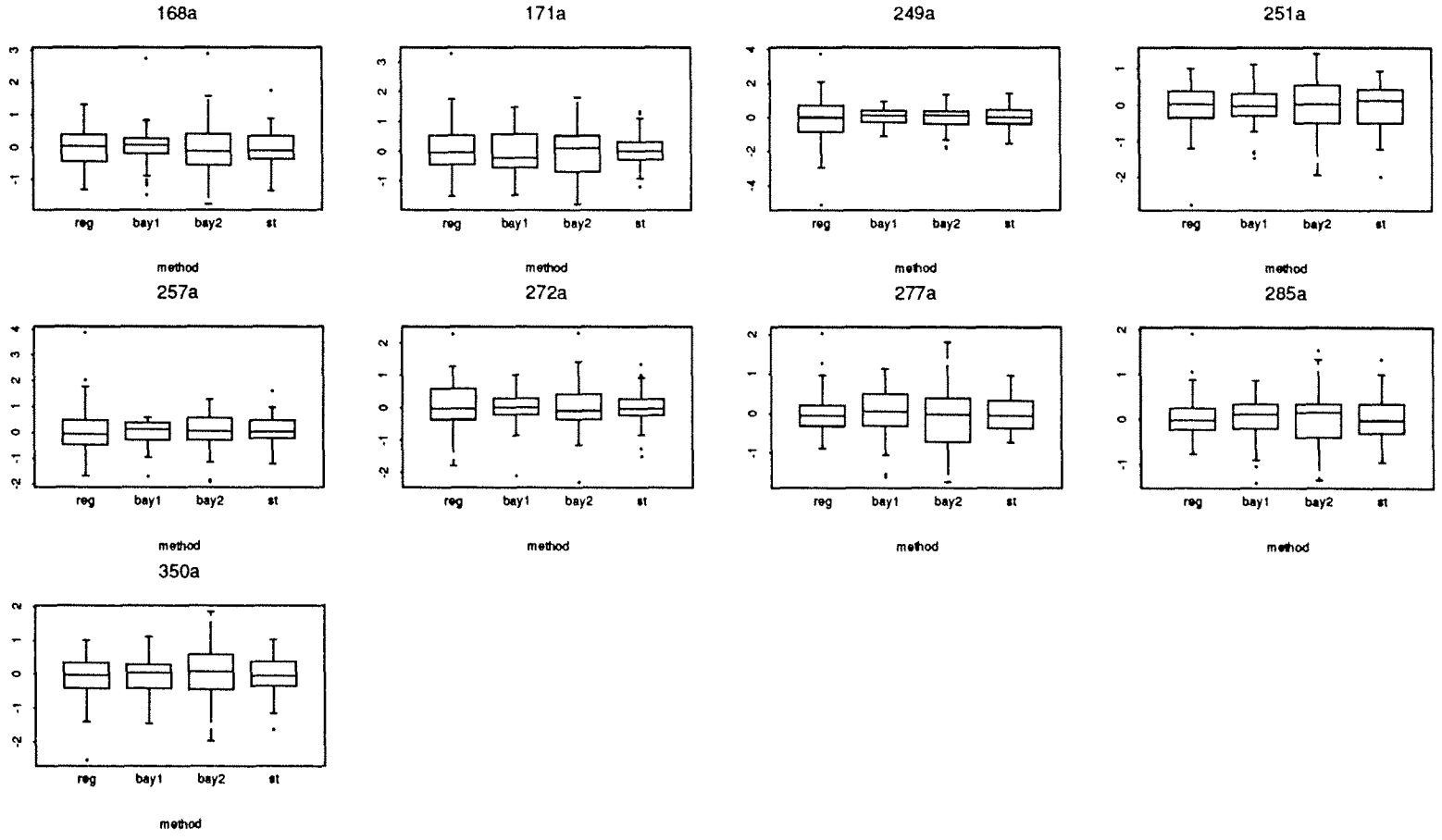
Legend:
reg=ordinary regression, bay1=regression using a Bayesian alternative (1) approach,
bay2= regression using a Bayesian alternative (2) approach, st= Stone's procedure

Figure A2.1(e): Continued



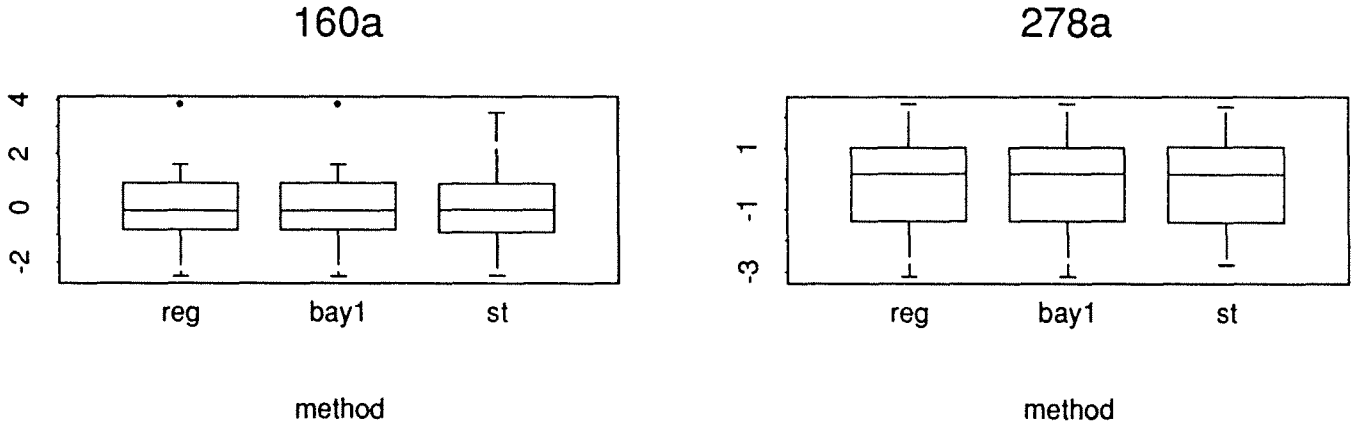
Legend:
reg=ordinary regression, bay1=regression using a Bayesian alternative (1) approach,
bay2= regression using a Bayesian alternative (2) approach, st= Stone's procedure

Figure A2.1(e): Continued



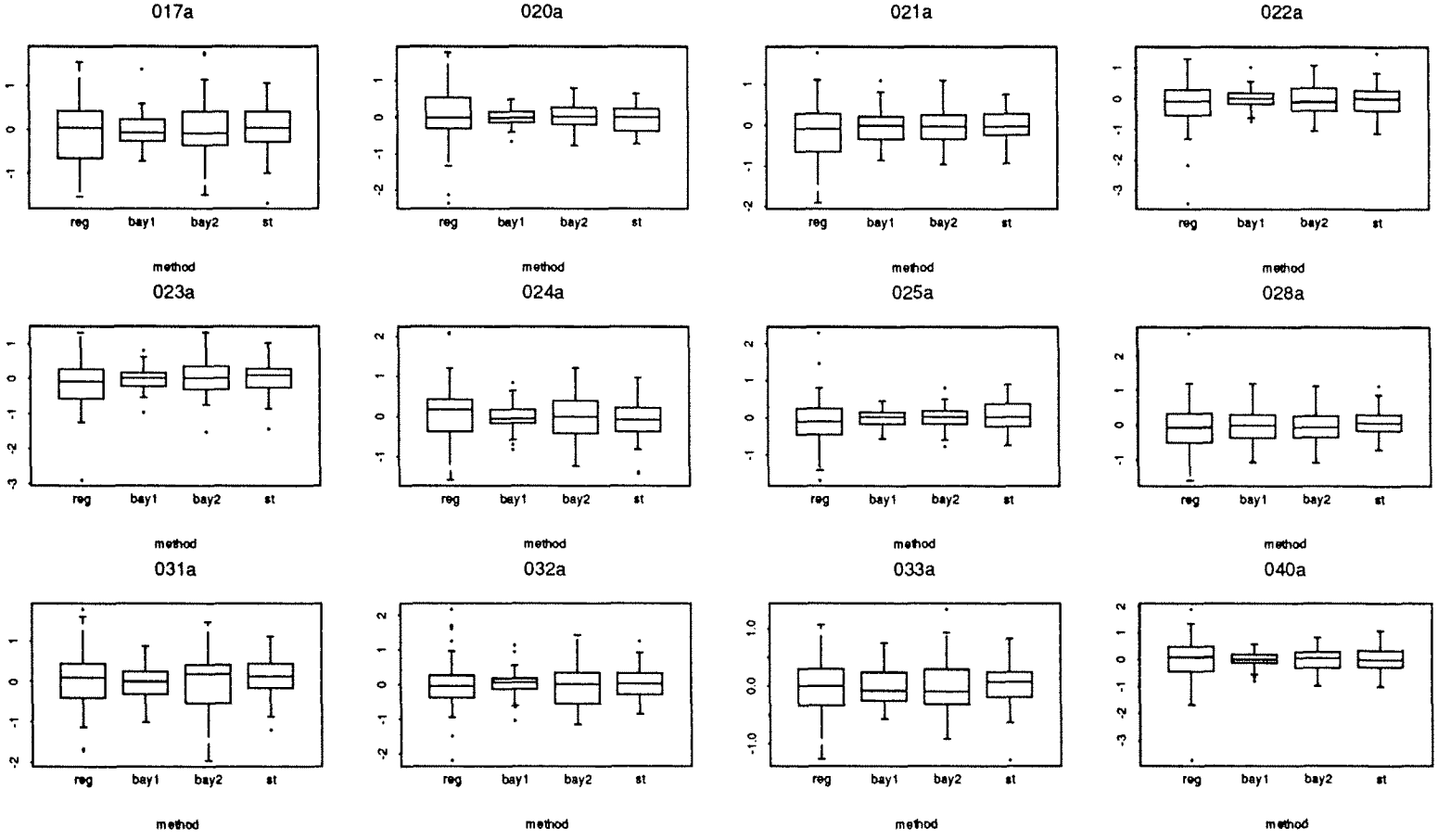
Legend:
reg=ordinary regression, bay1=regression using a Bayesian alternative (1) approach,
bay2= regression using a Bayesian alternative (2) approach, st= Stone's procedure

Figure 2.1(f): Boxplots of Prediction Errors for Cluster 6 of Hydrogen



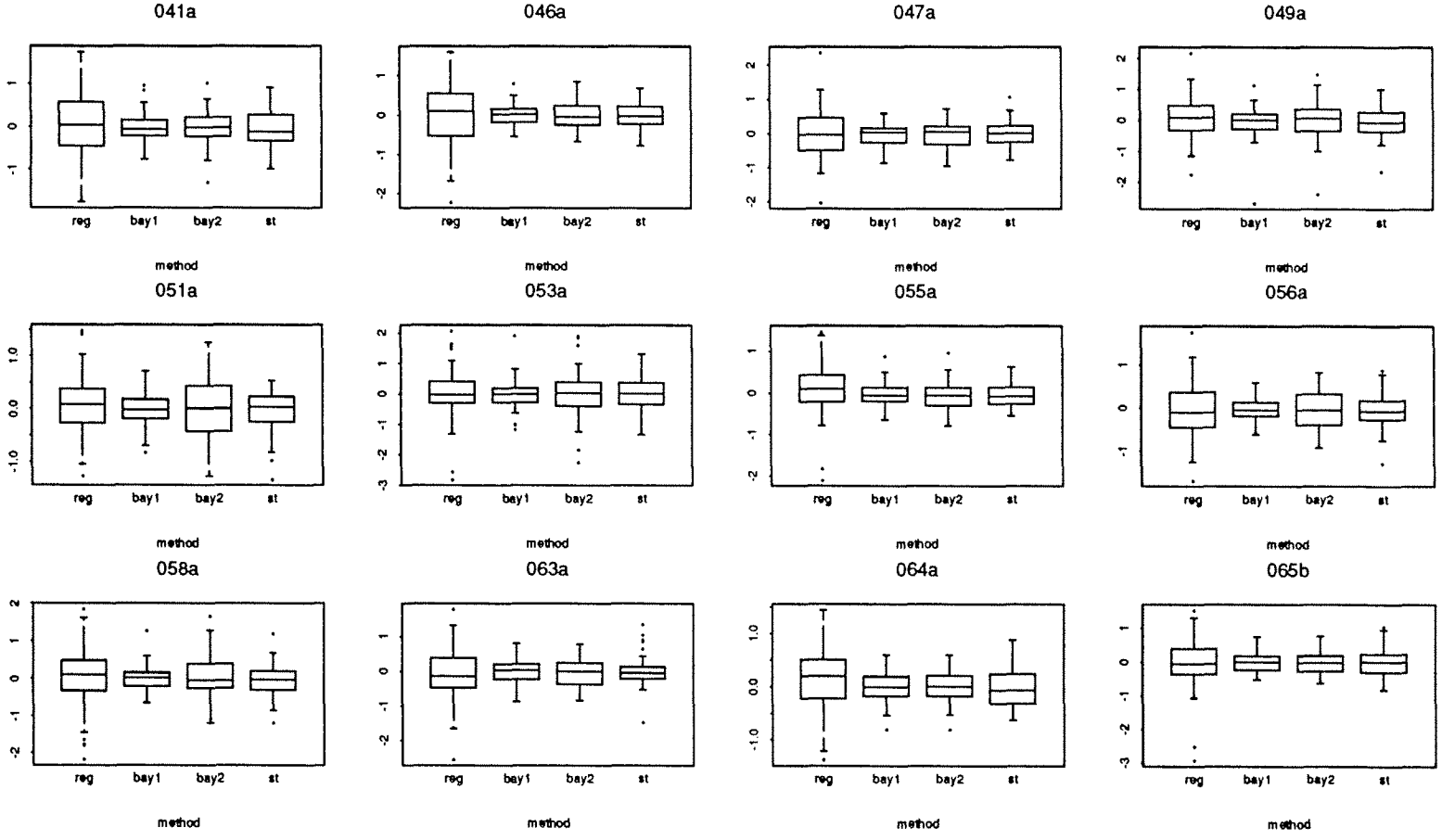
Legend:
reg=ordinary regression, bay1=regression using a Bayesian alternative (1) approach,
bay2= regression using a Bayesian alternative (2) approach, st= Stone's procedure

Figure A2.2(a): Boxplots of Prediction Errors for Cluster 1 of Sulfate



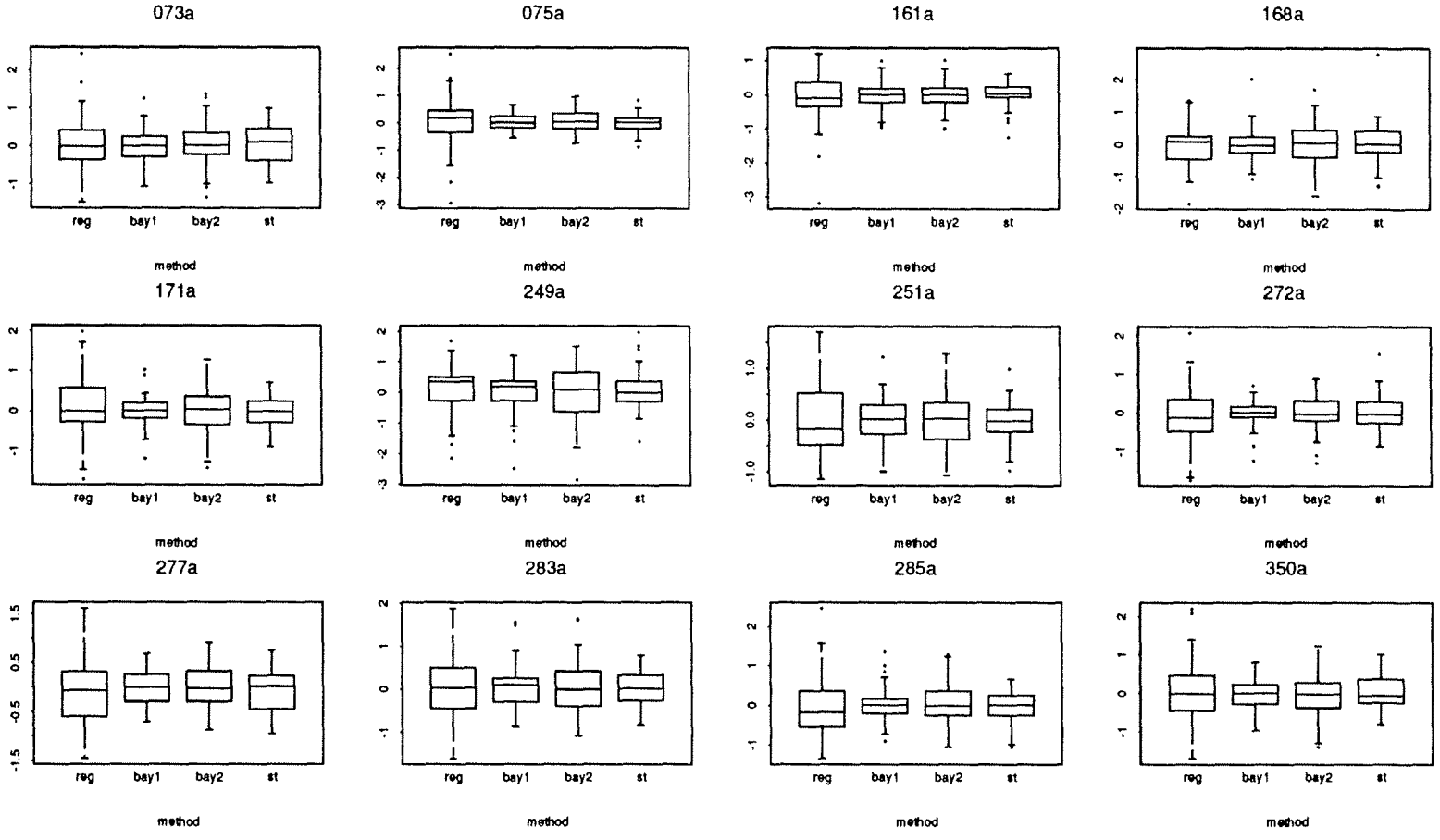
Legend:
reg=ordinary regression, bay1=regression using a Bayesian alternative (1) approach,
bay2= regression using a Bayesian alternative (2) approach, st= Stone's procedure

Figure A2.2(a): Continued



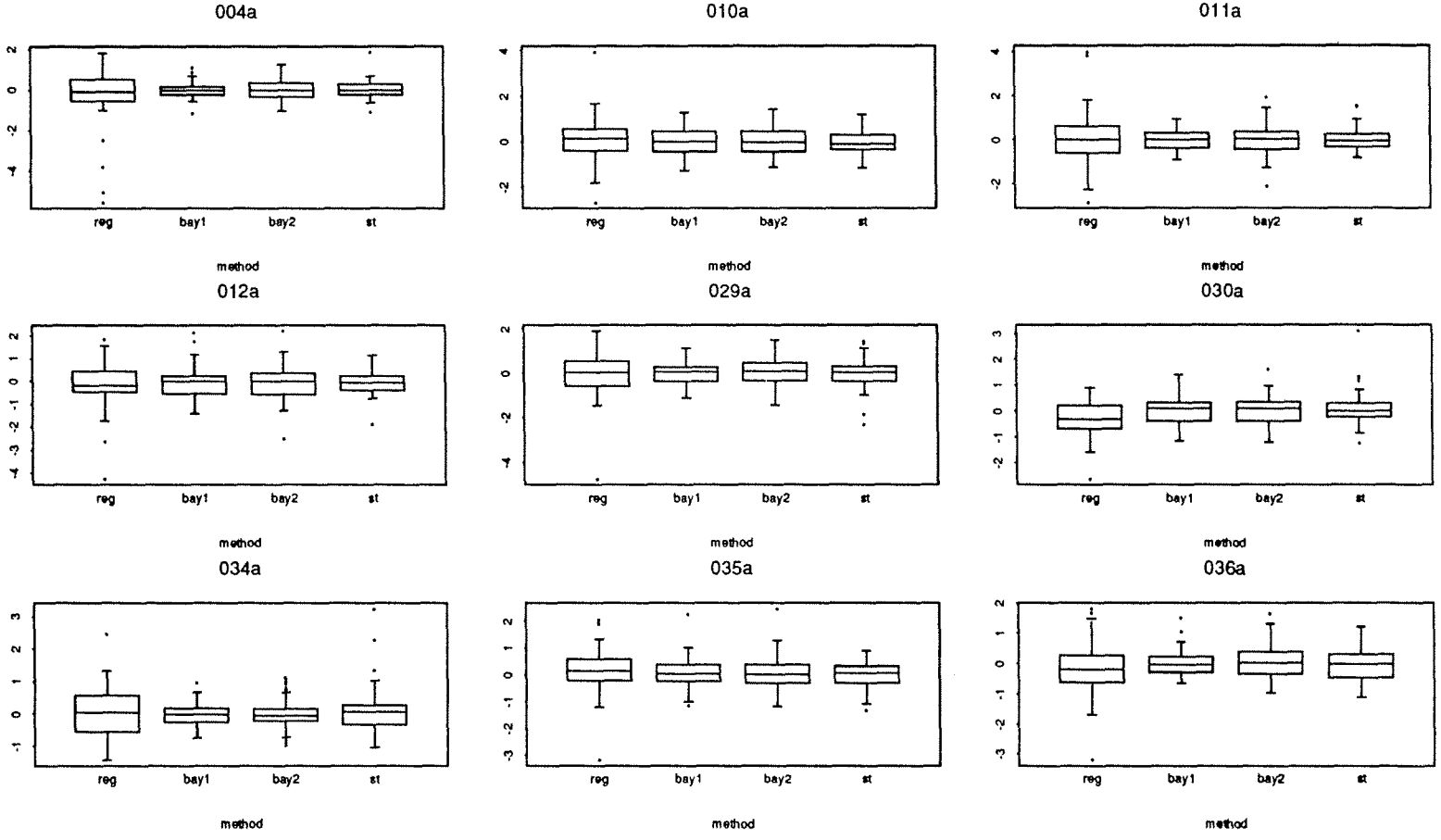
Legend:
reg=ordinary regression, bay1=regression using a Bayesian alternative (1) approach,
bay2= regression using a Bayesian alternative (2) approach, st= Stone's procedure

Figure A2.2(a): Continued



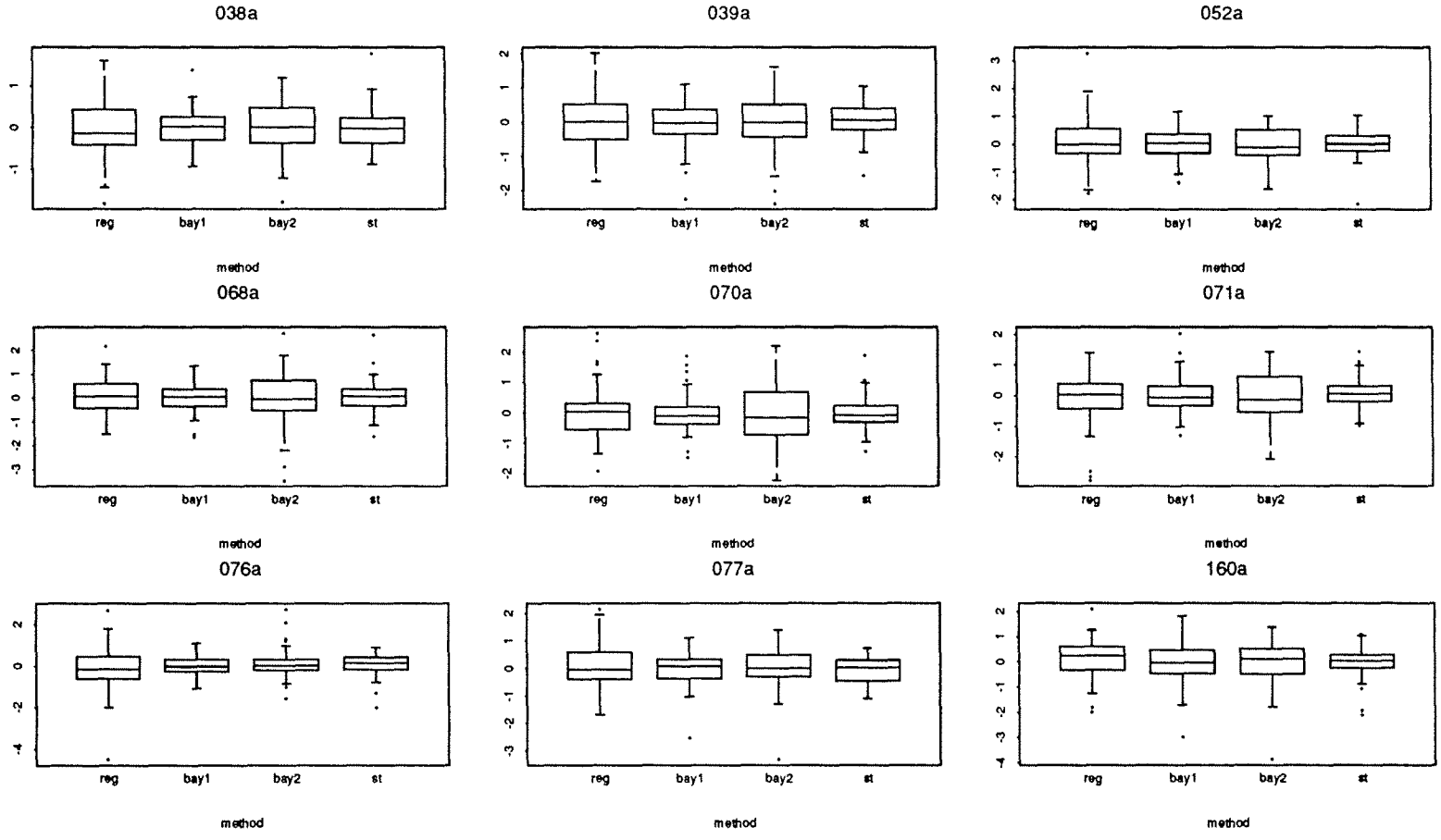
Legend:
reg=ordinary regression, bay1=regression using a Bayesian alternative (1) approach,
bay2= regression using a Bayesian alternative (2) approach, st= Stone's procedure

Figure A2.2(b): Boxplots of Prediction Errors for Cluster 2 of Sulfate



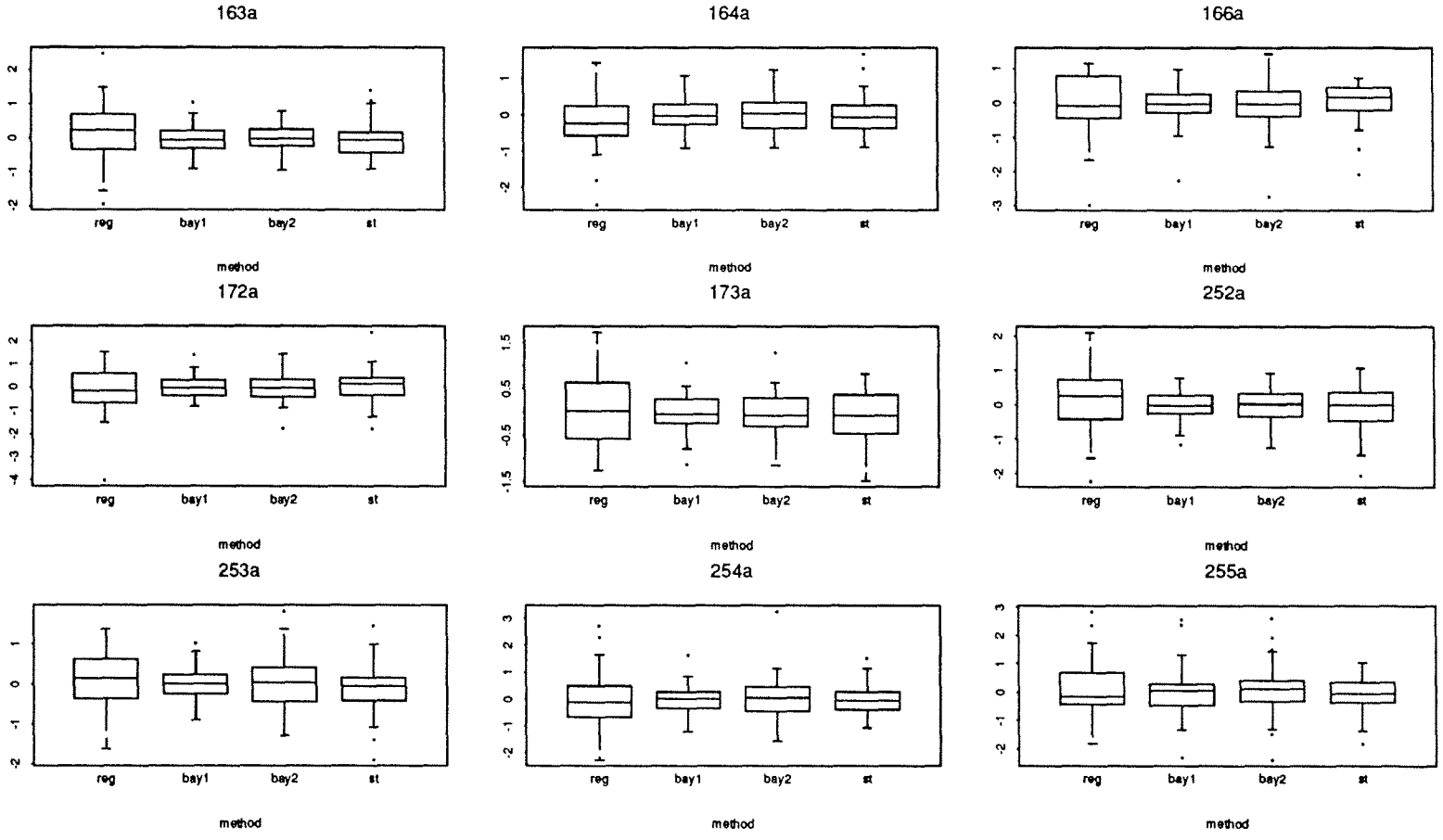
Legend:
reg=ordinary regression, bay1=regression using a Bayesian alternative (1) approach,
bay2= regression using a Bayesian alternative (2) approach, st= Stone's procedure

Figure A2.2(b): Continued



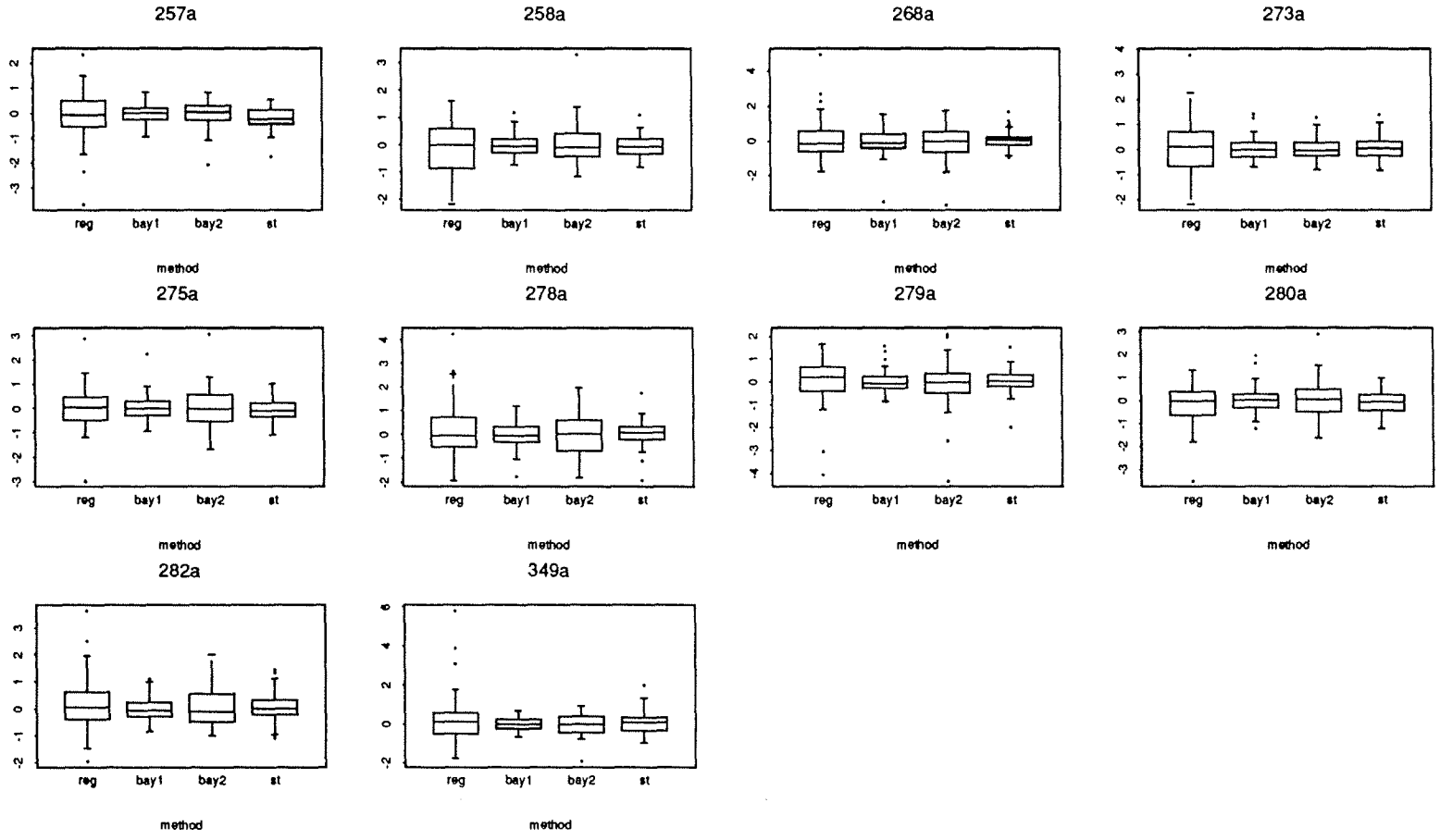
Legend:
reg=ordinary regression, bay1=regression using a Bayesian alternative (1) approach,
bay2= regression using a Bayesian alternative (2) approach, st= Stone's procedure

Figure A2.2(b): Continued



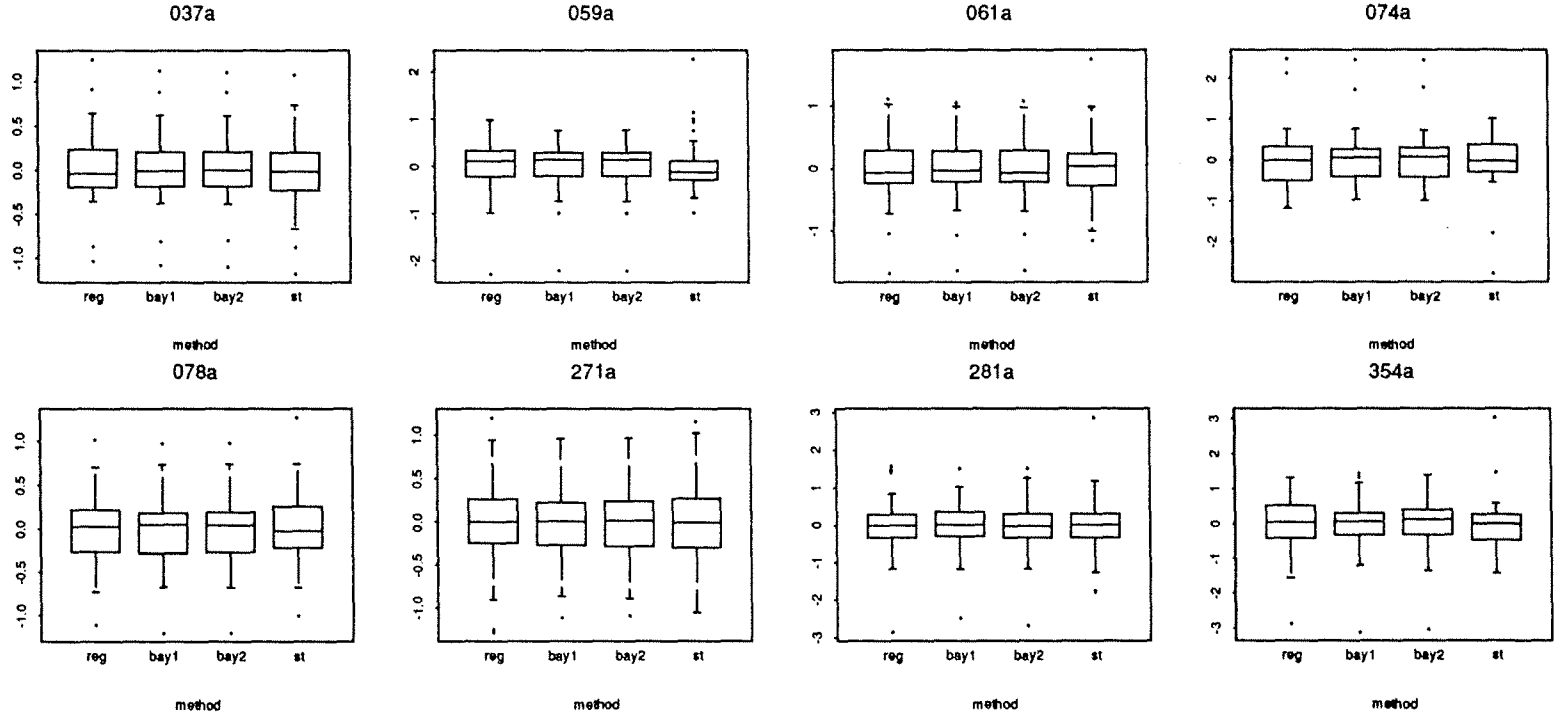
Legend:
reg=ordinary regression, bay1=regression using a Bayesian alternative (1) approach,
bay2= regression using a Bayesian alternative (2) approach, st= Stone's procedure

Figure A2.2(b): Continued



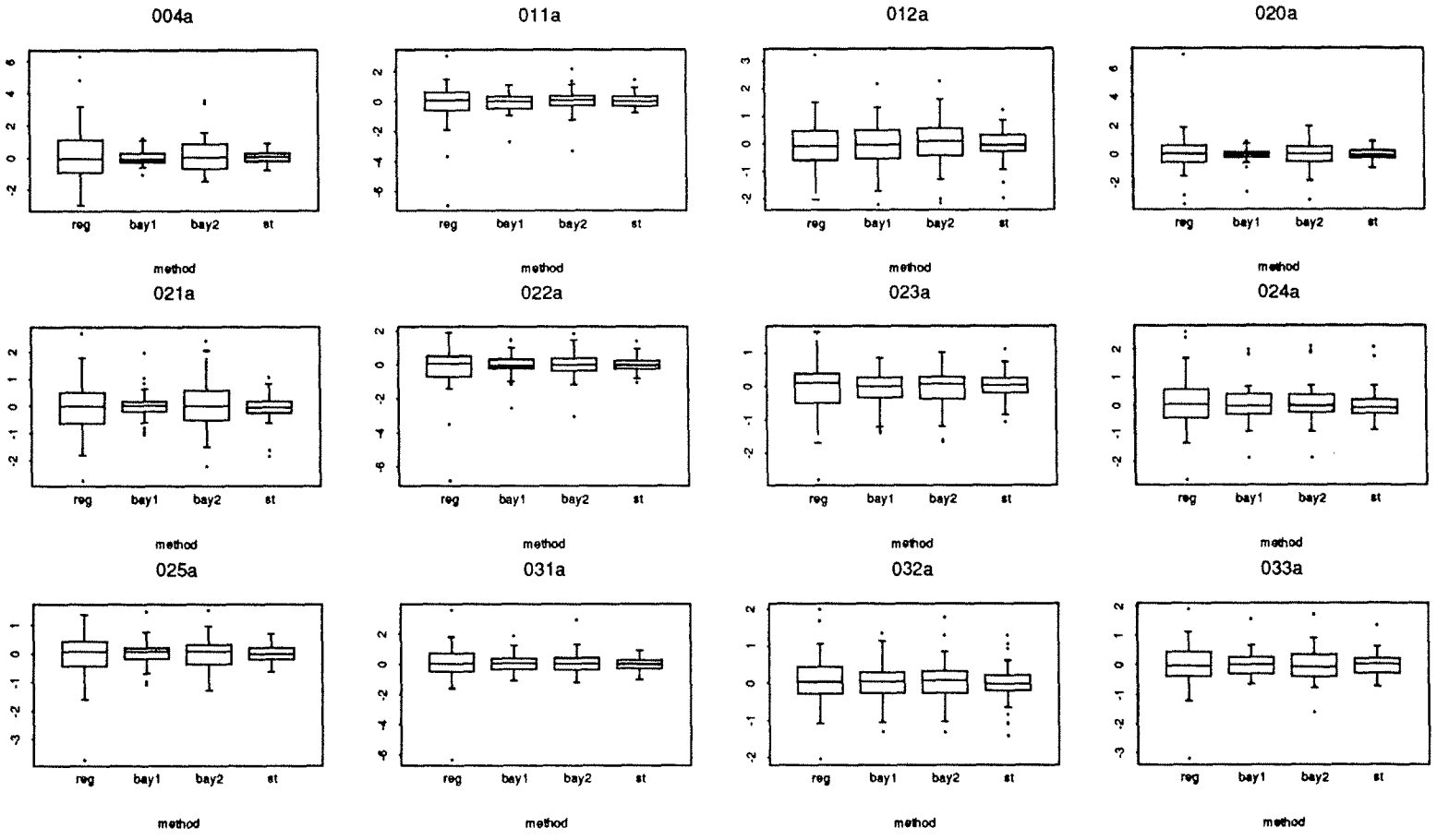
Legend:
reg=ordinary regression, bay1=regression using a Bayesian alternative (1) approach,
bay2= regression using a Bayesian alternative (2) approach, st= Stone's procedure

Figure A2.2(c): Boxplots of Prediction Errors for Cluster 3 of Sulfate



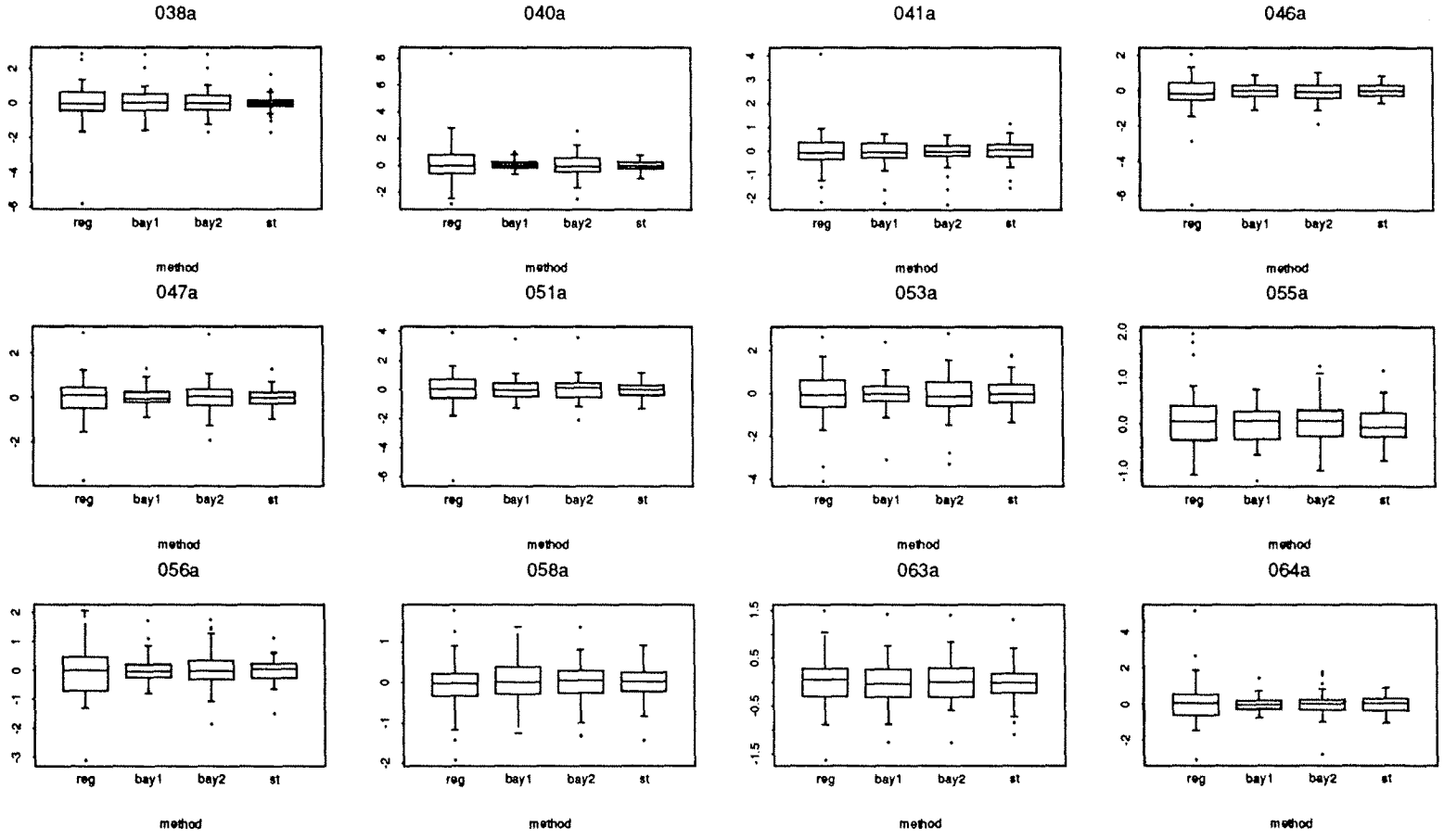
Legend:
reg=ordinary regression, bay1=regression using a Bayesian alternative (1) approach,
bay2= regression using a Bayesian alternative (2) approach, st= Stone's procedure

Figure A2.3(a): Boxplots of Prediction Errors for Cluster 1 of Nitrate



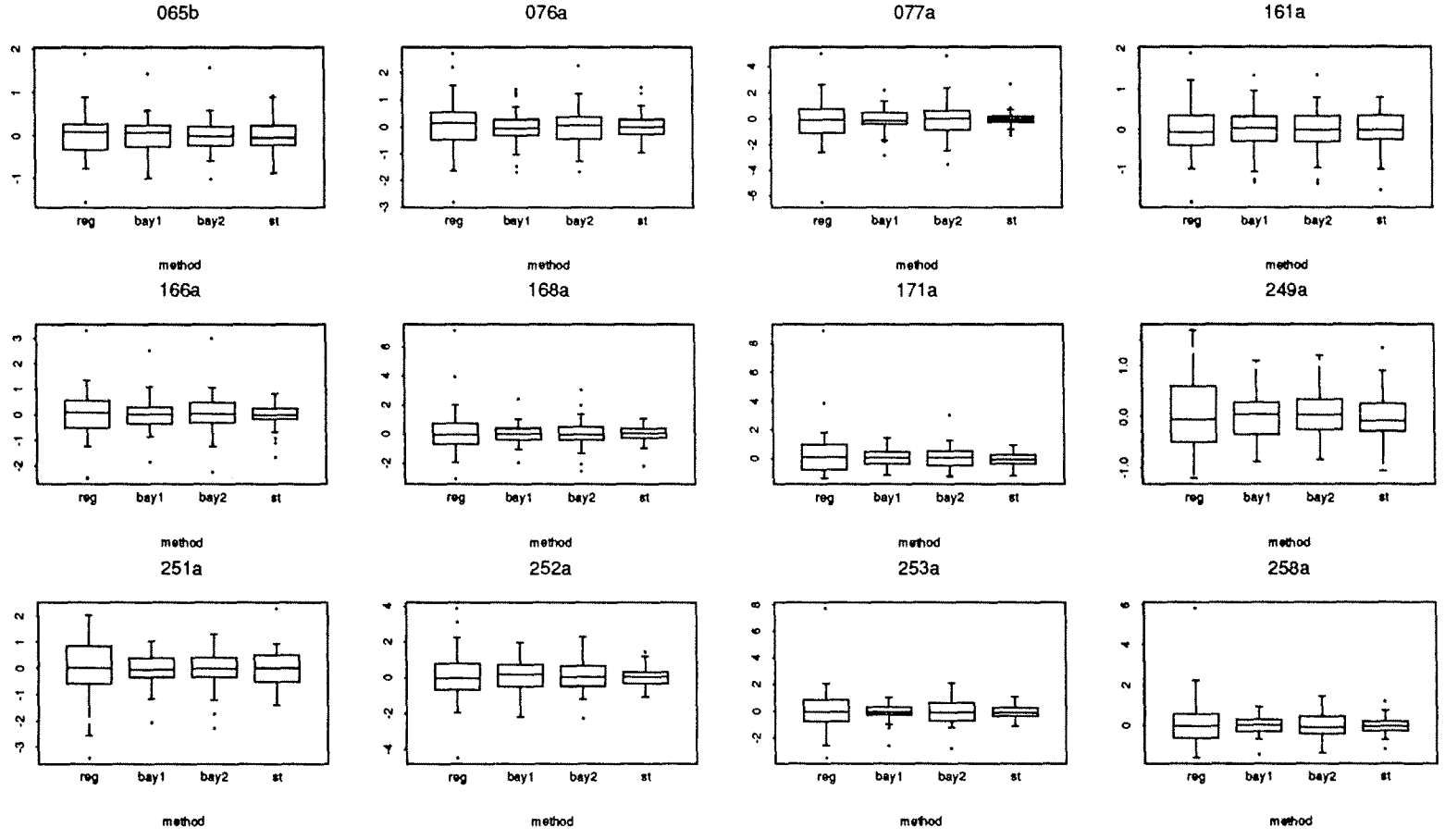
Legend:
reg=ordinary regression, bay1=regression using a Bayesian alternative (1) approach,
bay2= regression using a Bayesian alternative (2) approach, st= Stone's procedure

Figure A2.3(a): Continued



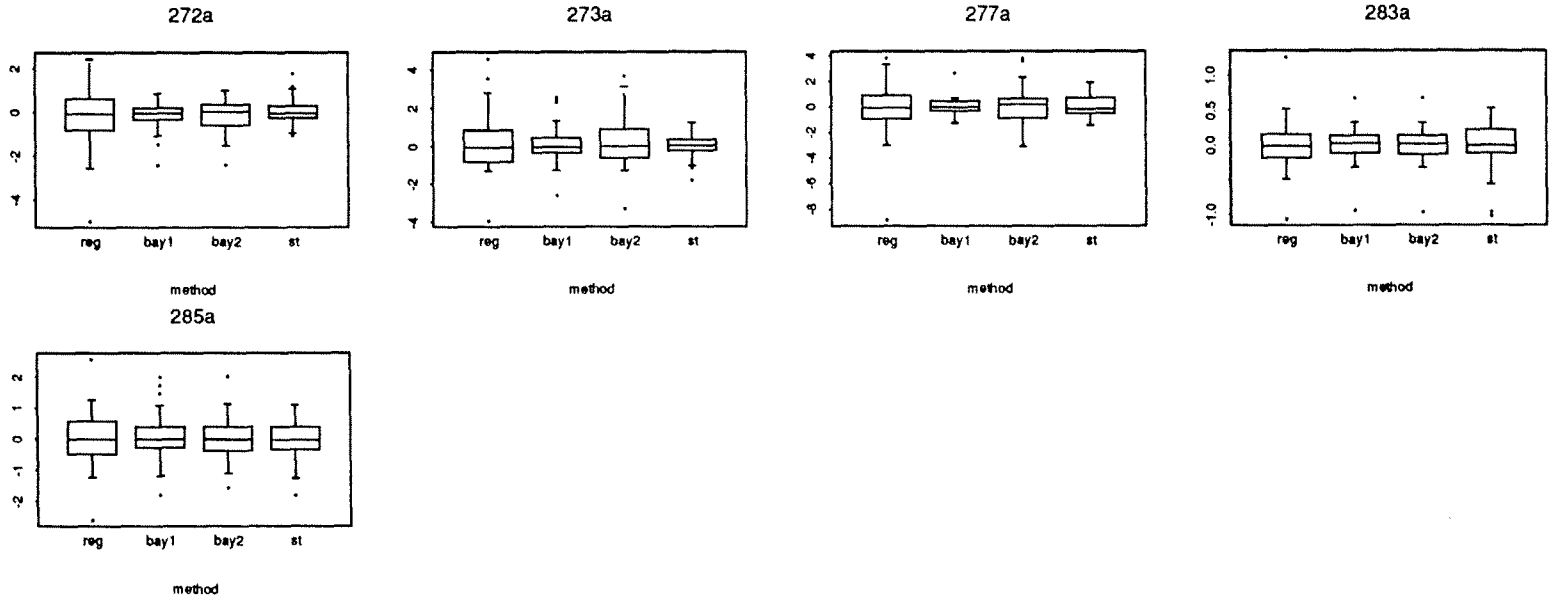
Legend:
reg=ordinary regression, bay1=regression using a Bayesian alternative (1) approach,
bay2= regression using a Bayesian alternative (2) approach, st= Stone's procedure

Figure A2.3(a): Continued



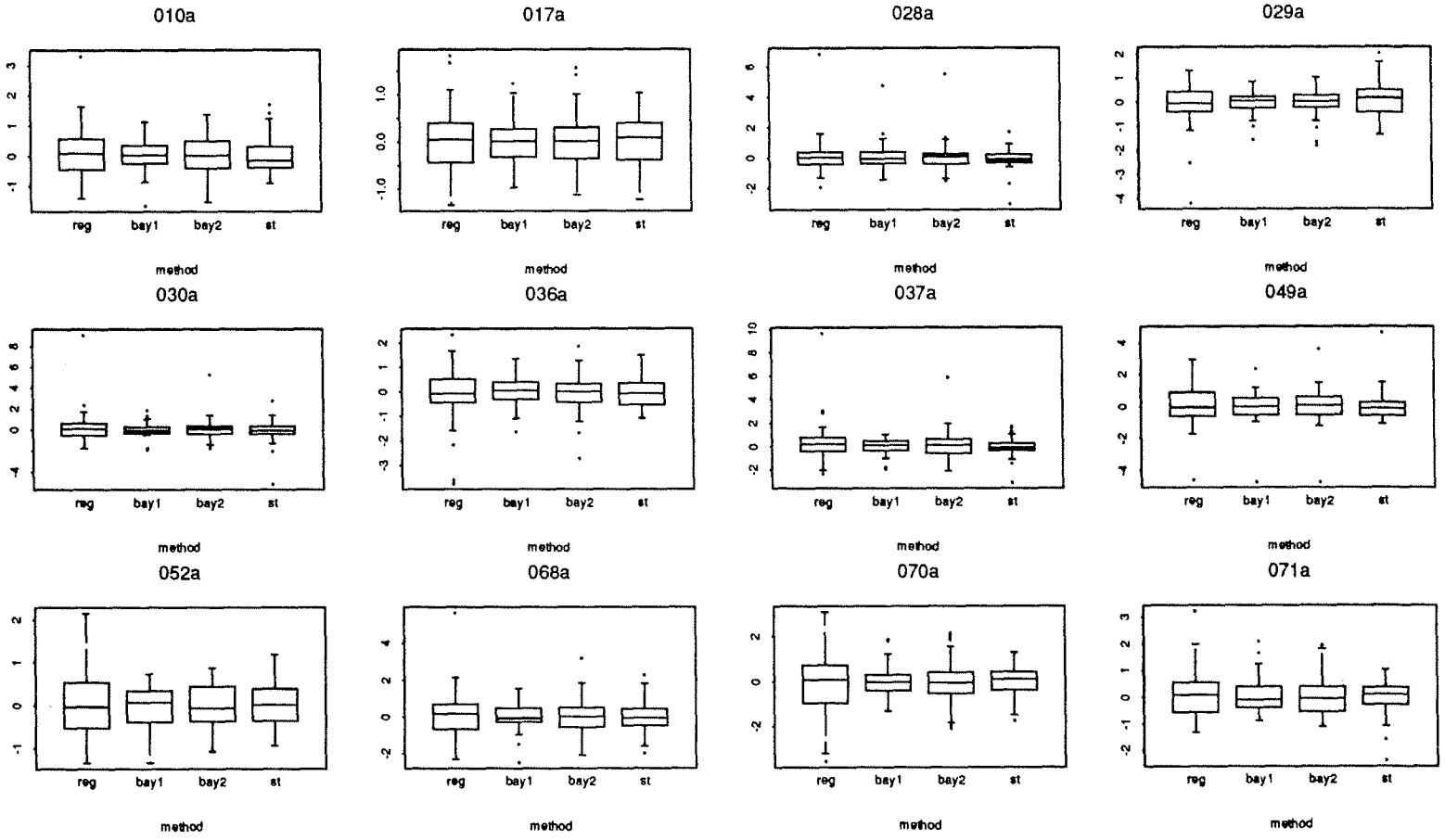
Legend:
reg=ordinary regression, bay1=regression using a Bayesian alternative (1) approach,
bay2= regression using a Bayesian alternative (2) approach, st= Stone's procedure

Figure A2.3(a): Continued



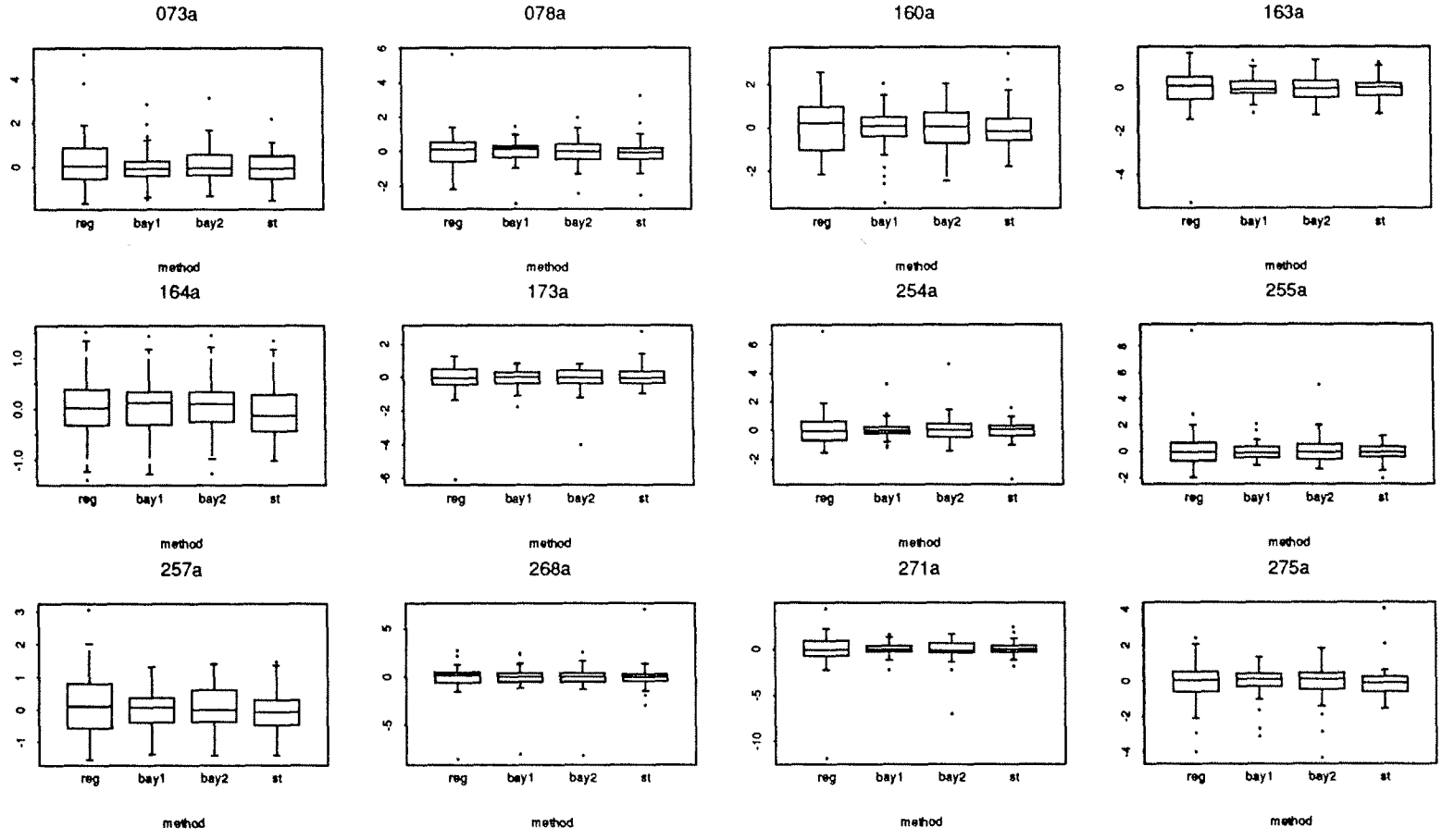
Legend:
reg=ordinary regression, bay1=regression using a Bayesian alternative (1) approach,
bay2= regression using a Bayesian alternative (2) approach, st= Stone's procedure

Figure A2.3(b): Boxplots of Prediction Errors for Cluster 2 of Nitrate



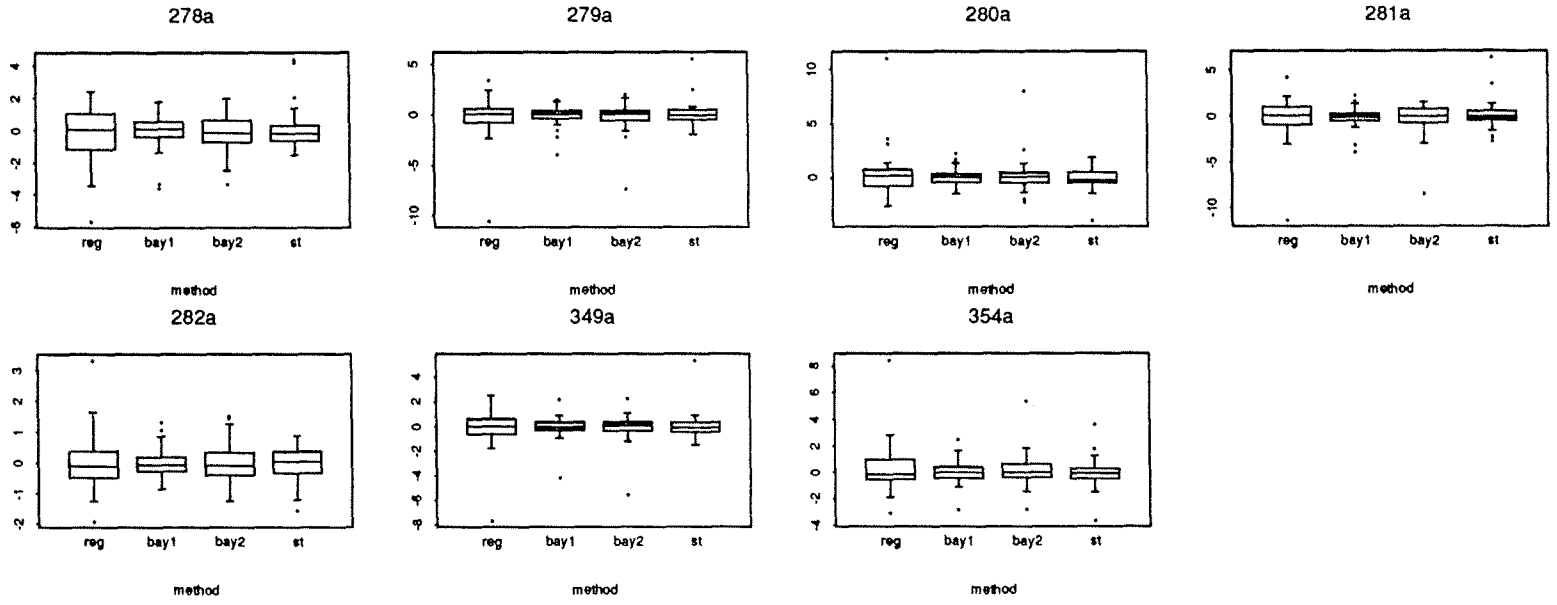
Legend:
reg=ordinary regression, bay1=regression using a Bayesian alternative (1) approach,
bay2= regression using a Bayesian alternative (2) approach, st= Stone's procedure

Figure A2.3(b): Continued



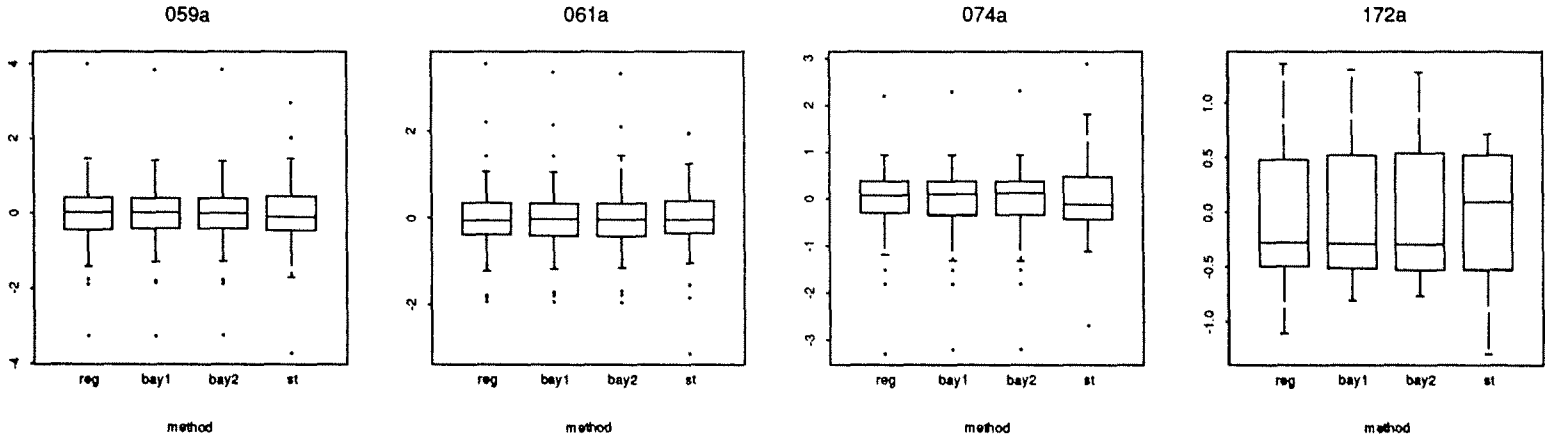
Legend:
reg=ordinary regression, bay1=regression using a Bayesian alternative (1) approach,
bay2= regression using a Bayesian alternative (2) approach, st= Stone's procedure

Figure A2.3(b): Continued



Legend:
reg=ordinary regression, bay1=regression using a Bayesian alternative (1) approach,
bay2= regression using a Bayesian alternative (2) approach, st= Stone's procedure

Figure A2.3(c): Boxplots of Prediction Errors for Cluster 3 of Nitrate



Legend:
reg=ordinary regression, bay1=regression using a Bayesian alternative (1) approach,
bay2= regression using a Bayesian alternative (2) approach, st= Stone's procedure

Figure A3.1(a): Relative Measure of Agreement for Hydrogen

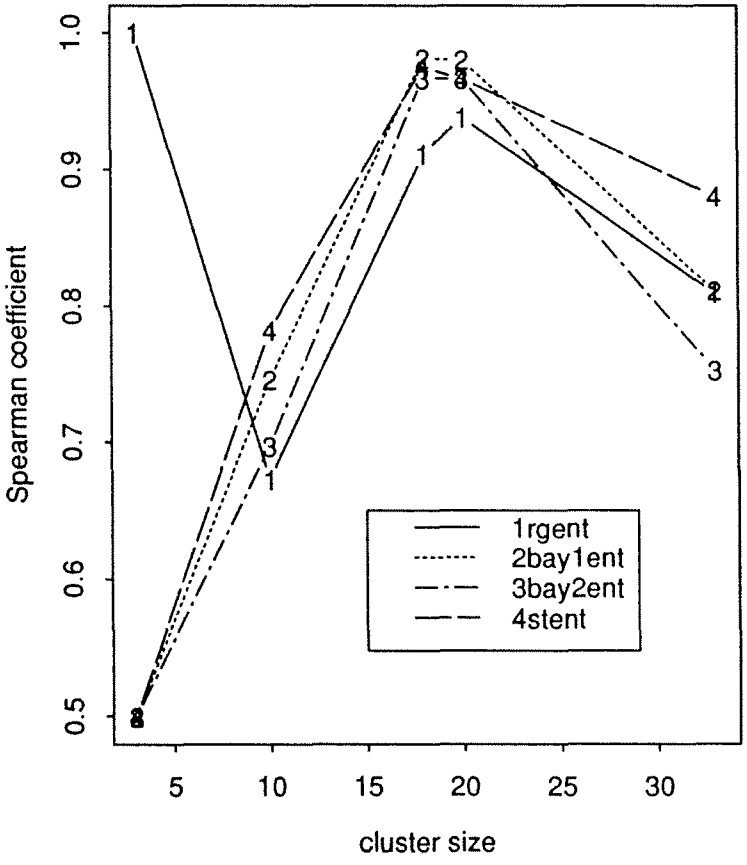
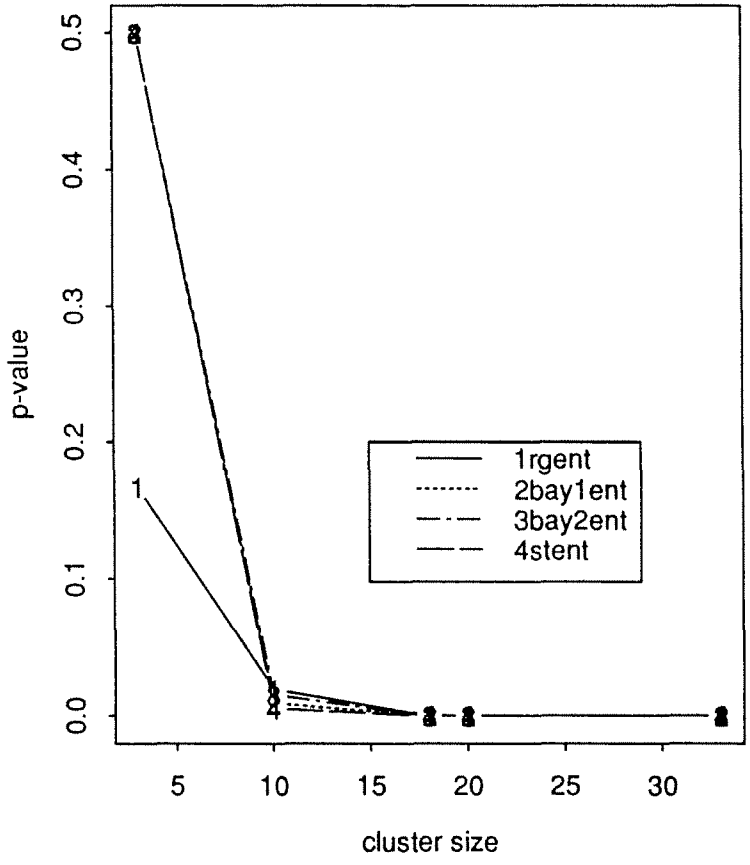


Figure A3.1(b): P-value for Tests of Agreement for Hydrogen



Legend:
rgent = regression with entropy, bay1ent= Bayesian alternative (1) with entropy
bay1ent= Bayesian alternative (2) with entropy, stent = Stone's procedure with entropy

Figure A3.2(a): Relative Measure of Agreement for Sulfate

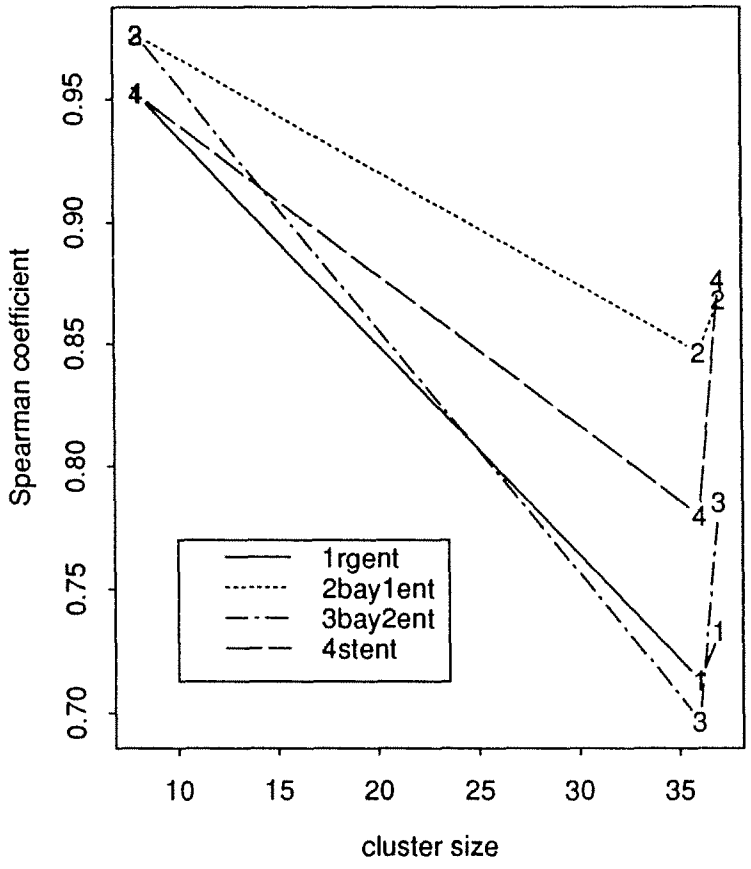
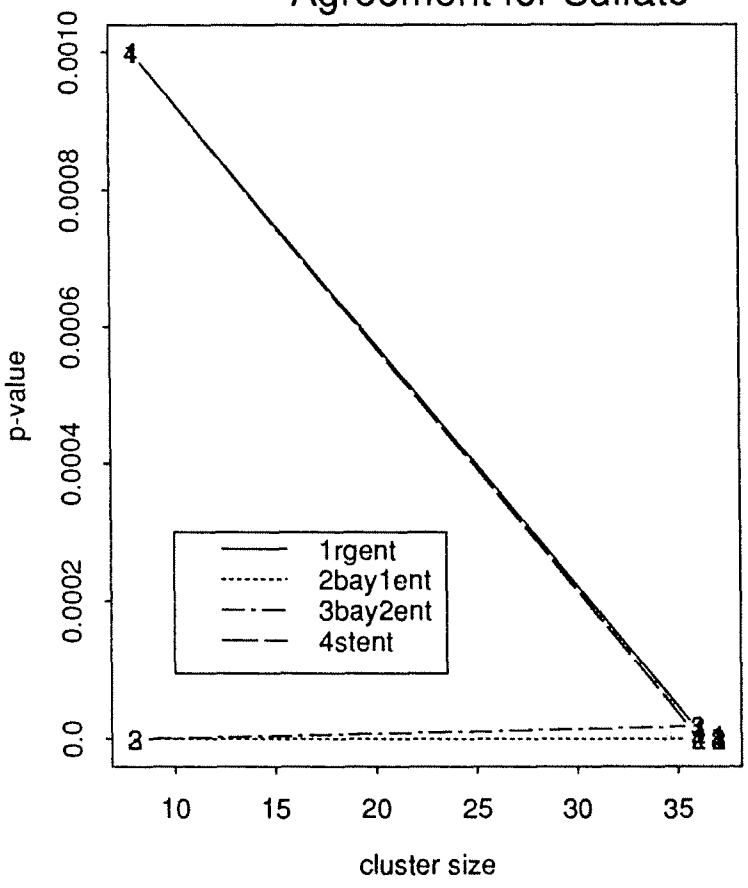


Figure A3.2(b): P-value for Tests of Agreement for Sulfate



Legend:
rgent = regression with entropy, bay1ent= Bayesian alternative (1) with entropy
bay1ent= Bayesian alternative (2) with entropy, stent = Stone's procedure with entropy

Figure A3.3(a): Relative Measure of Agreement for Nitrate

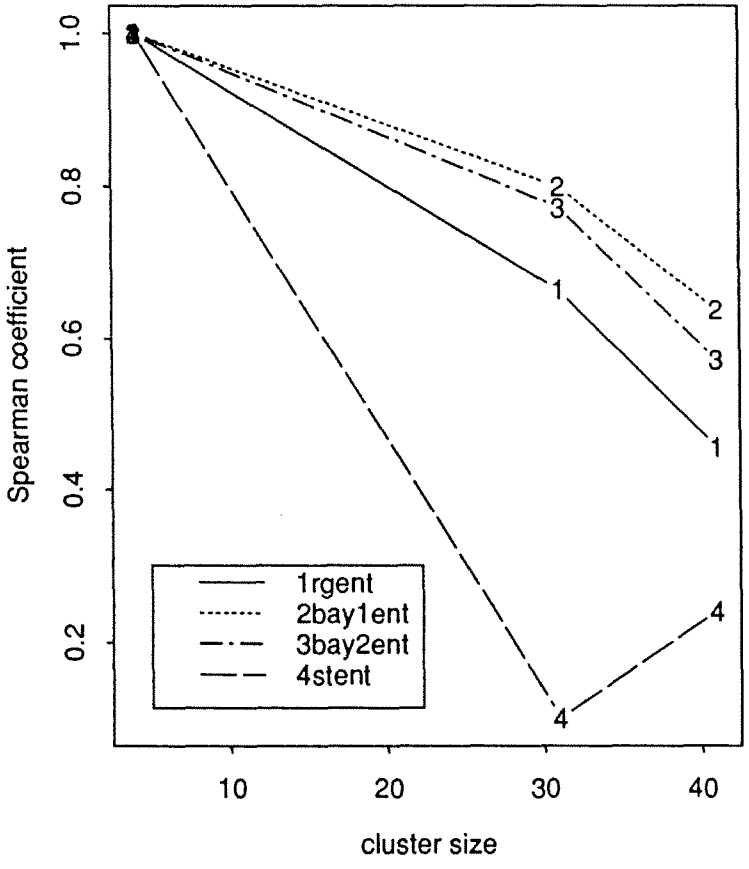
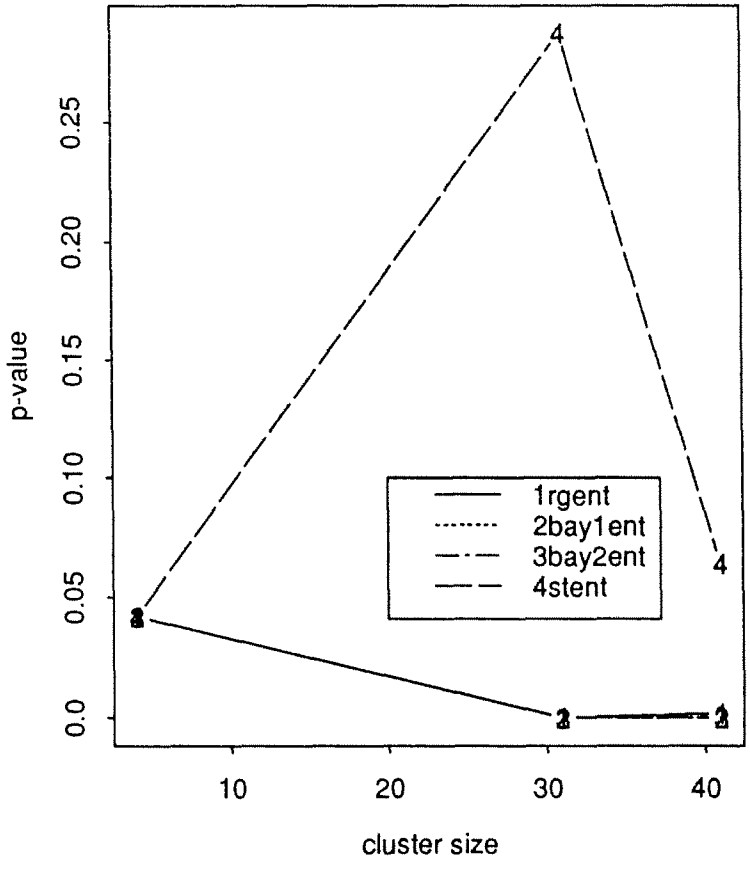


Figure A3.3(b): P-value for Tests of Agreement for Nitrate



Legend:
rgent = regression with entropy, bay1ent= Bayesian alternative (1) with entropy
bay1ent= Bayesian alternative (2) with entropy, stent = Stone's procedure with entropy

Average Prediction Errors for Hydrogen in Ascending Order

Figure A4.1(a): Cluster 1

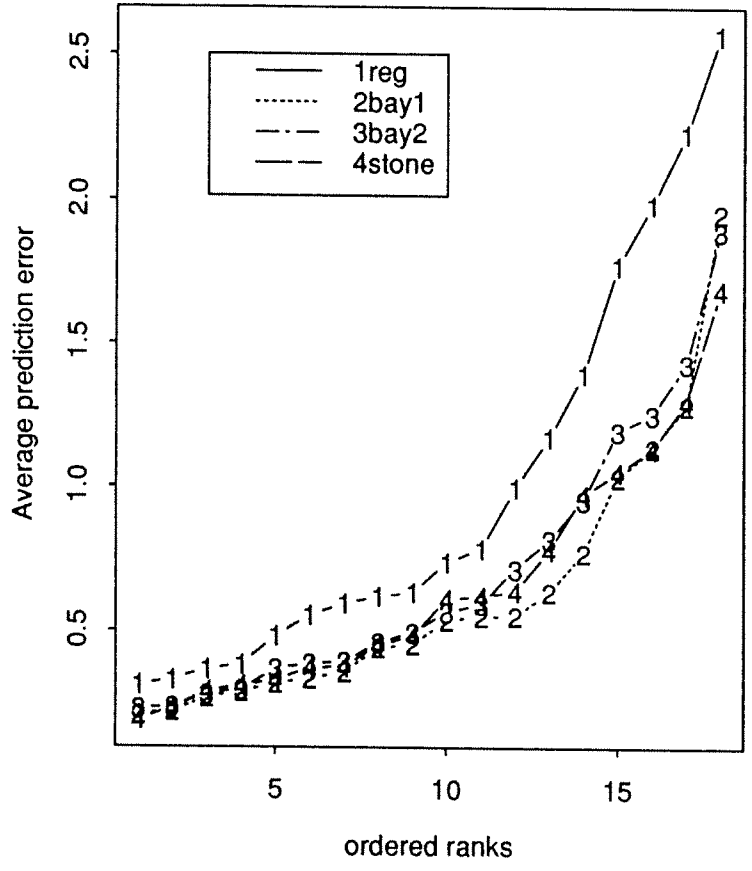
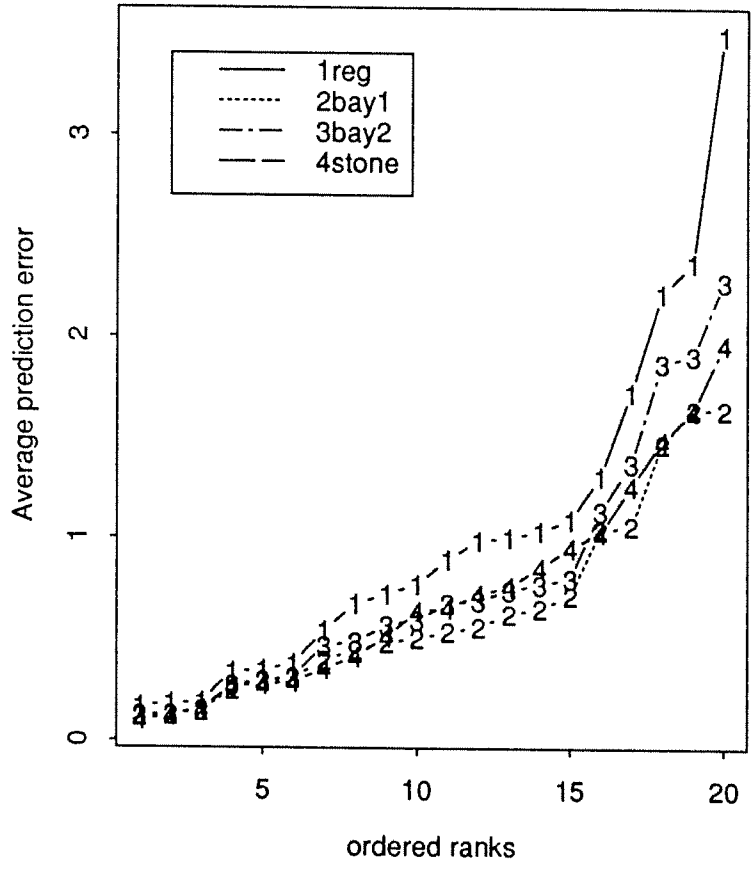


Figure A4.1(b): Cluster 2



Legend:
reg=regression,bay1= Bayesian (alternative1), bay2= Bayesian (alternative2)
Stone = Stone's procedure

Average Prediction Errors for Hydrogen in Ascending Order

Figure A4.1(c): Cluster 3

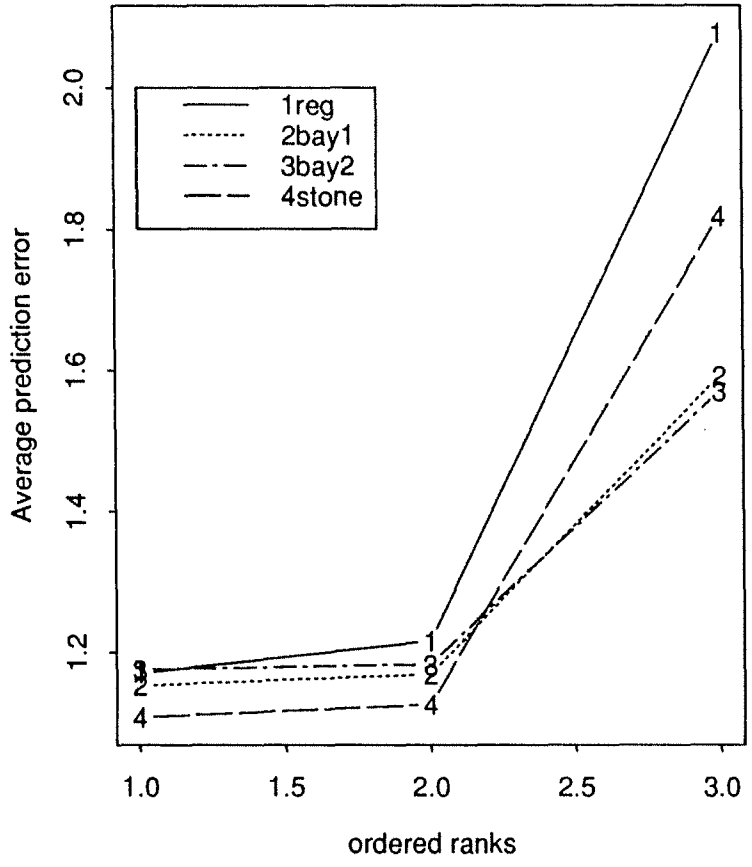
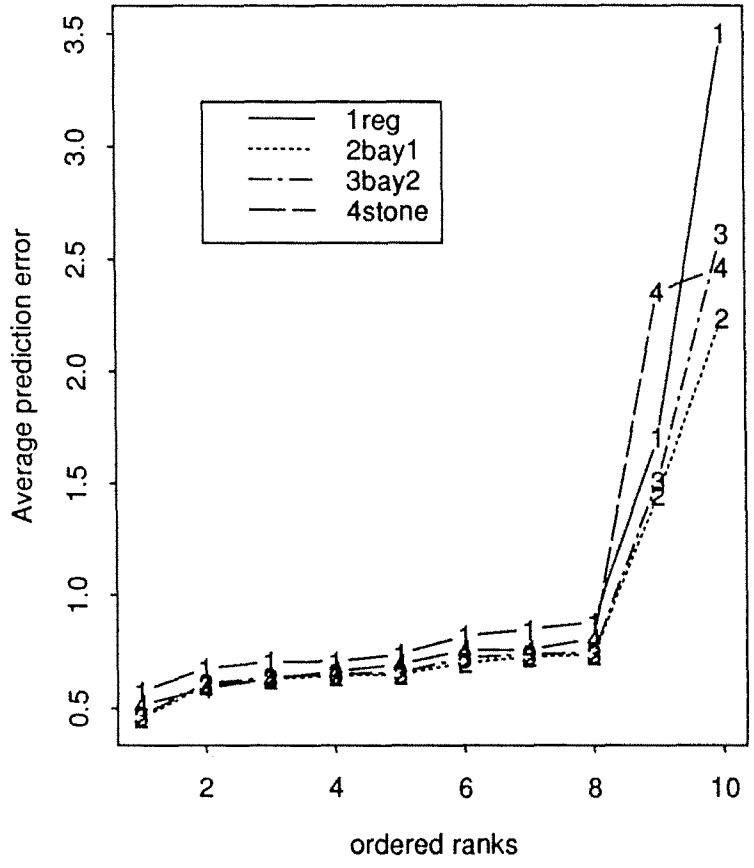


Figure A4.1(d): Cluster 4



Legend:
reg=regression, bay1= Bayesian alternative (1), bay2= Bayesian alternative (2)
Stone = Stone's procedure

Average Prediction Errors for Hydrogen in Ascending Order

Figure A4.1(e): Cluster 5

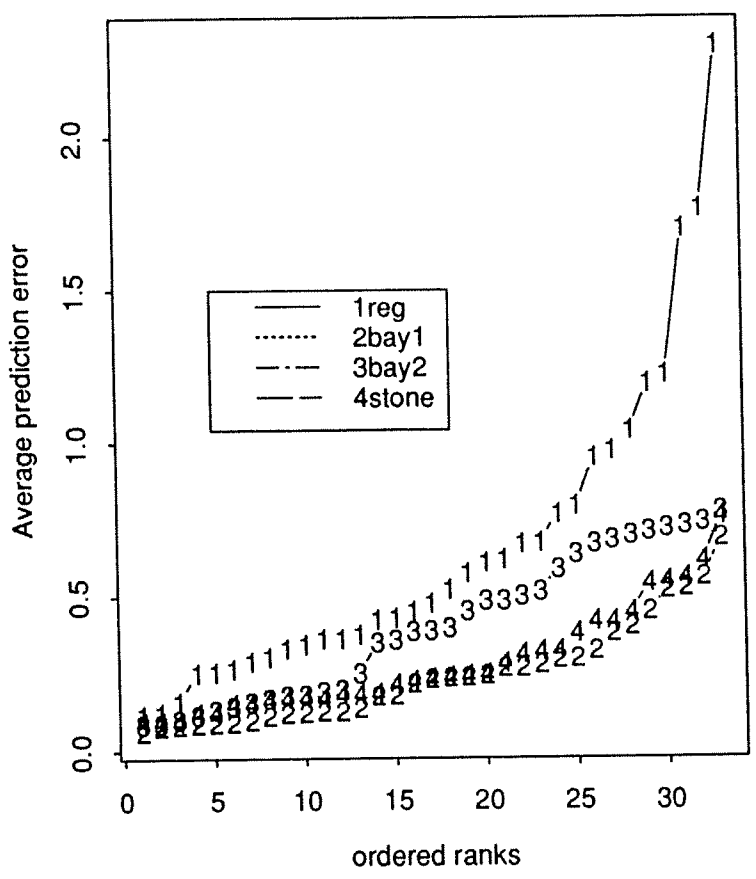
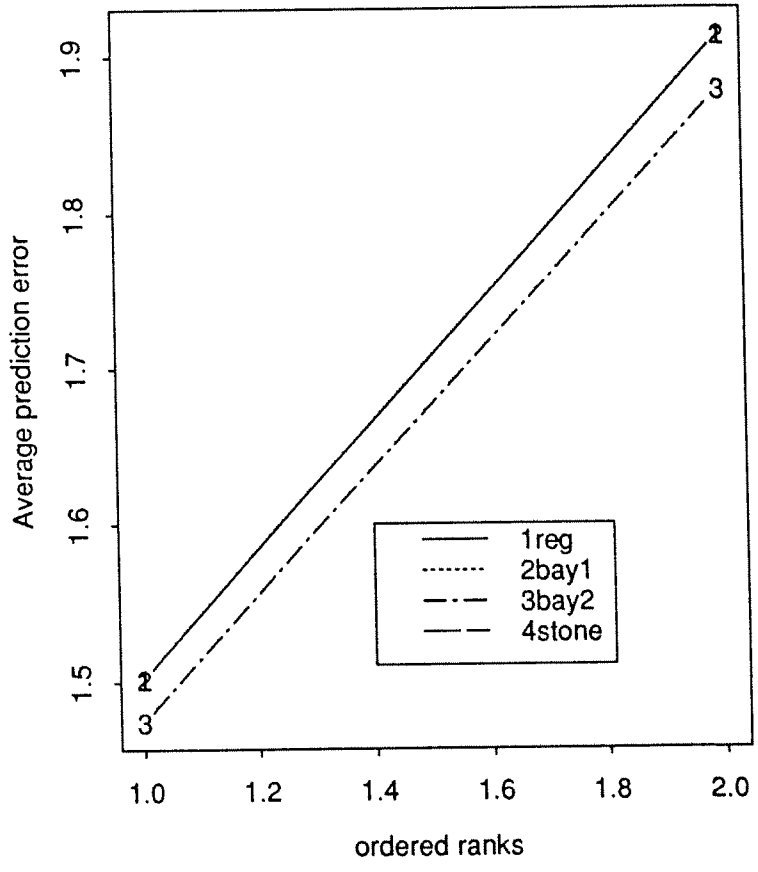


Figure A4.1(f): Cluster 6



Legend:
reg=regression, bay1= Bayesian alternative (1), bay2= Bayesian alternative (2)
Stone = Stone's procedure

Average Prediction Errors for Sulfate in Ascending Order

Figure A4.2(a): Cluster 1

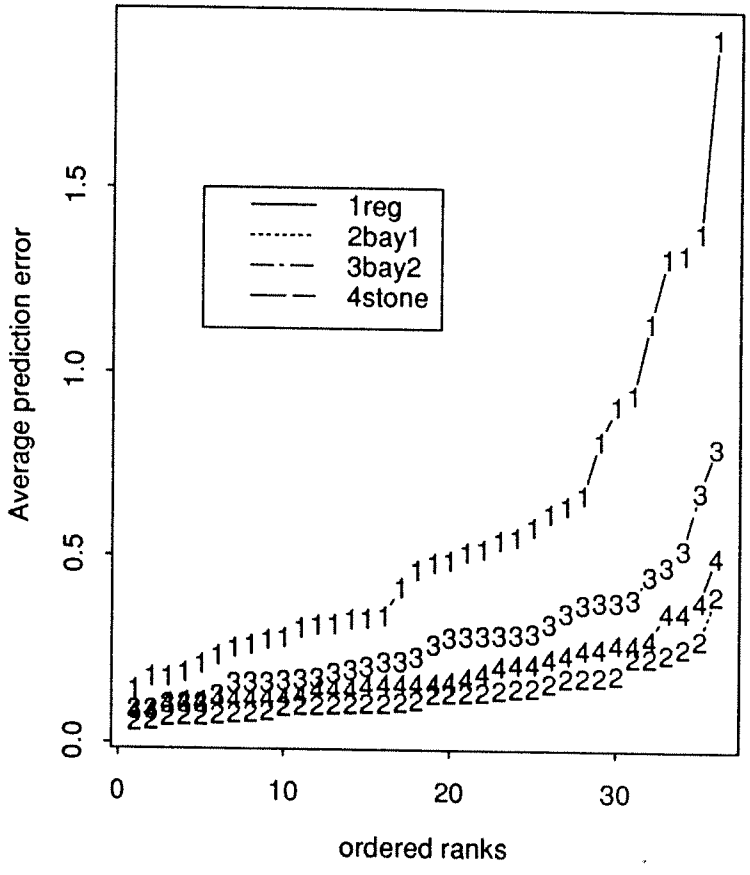
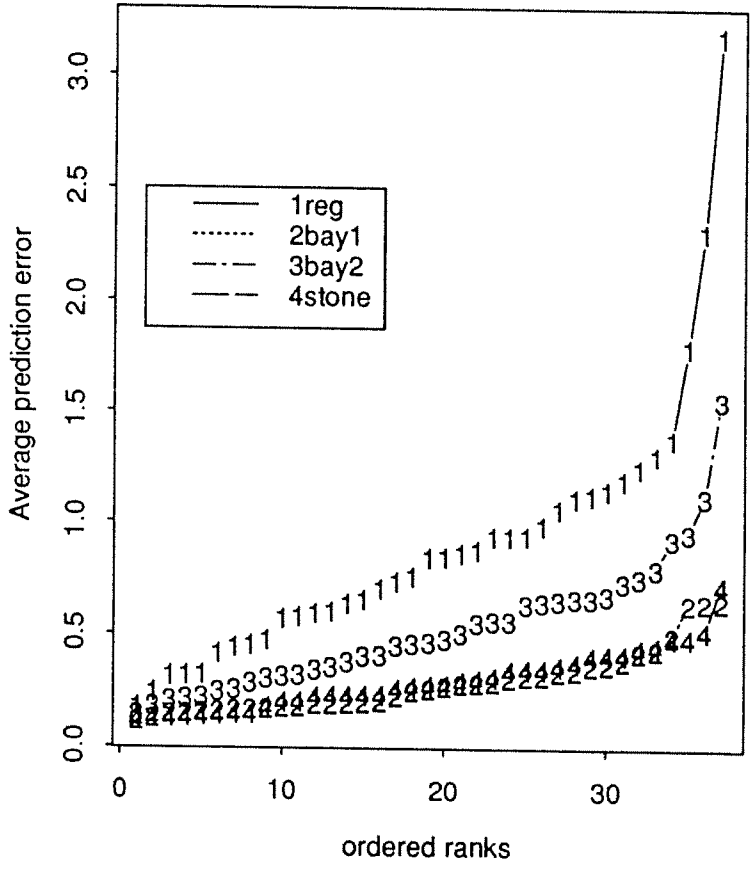


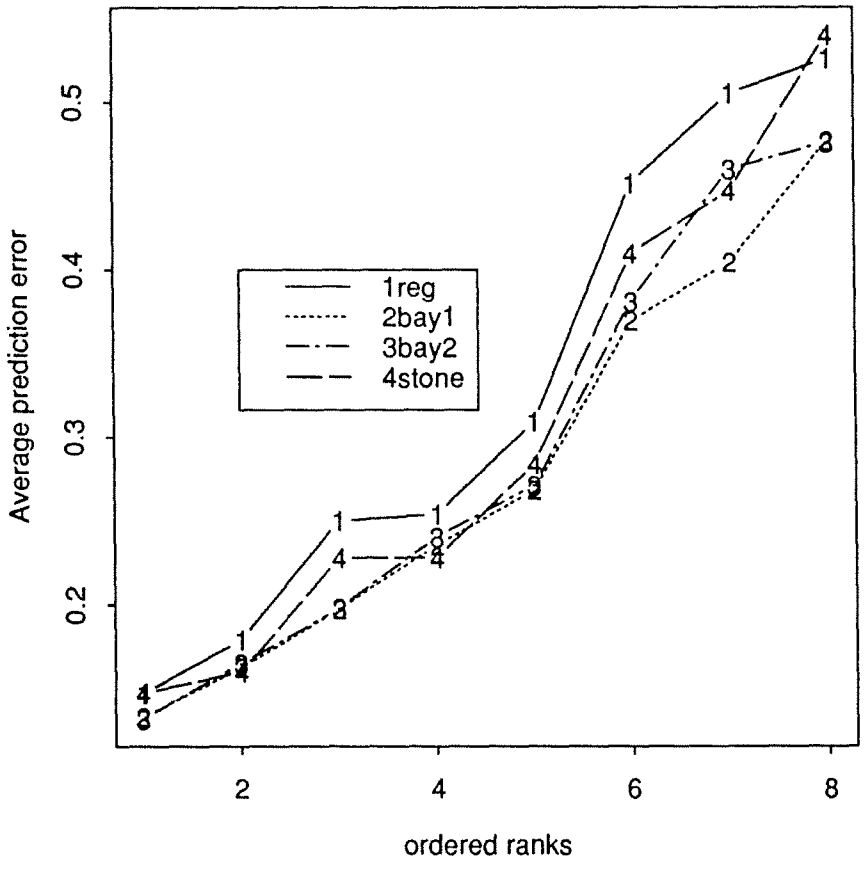
Figure A4.2(b): Cluster 2



Legend:
reg=regression, bay1= Bayesian alternative (1), bay2= Bayesian alternative (2)
Stone = Stone's procedure

Average Prediction Errors for Sulfate in Ascending Order

Figure A4.2(c): Cluster 3



Legend:
reg=regression, bay1= Bayesian alternative (1), bay2= Bayesian alternative (2)
Stone = Stone's procedure

Average Prediction Errors for Nitrate in Ascending Order

Figure A4.3(a): Cluster 1

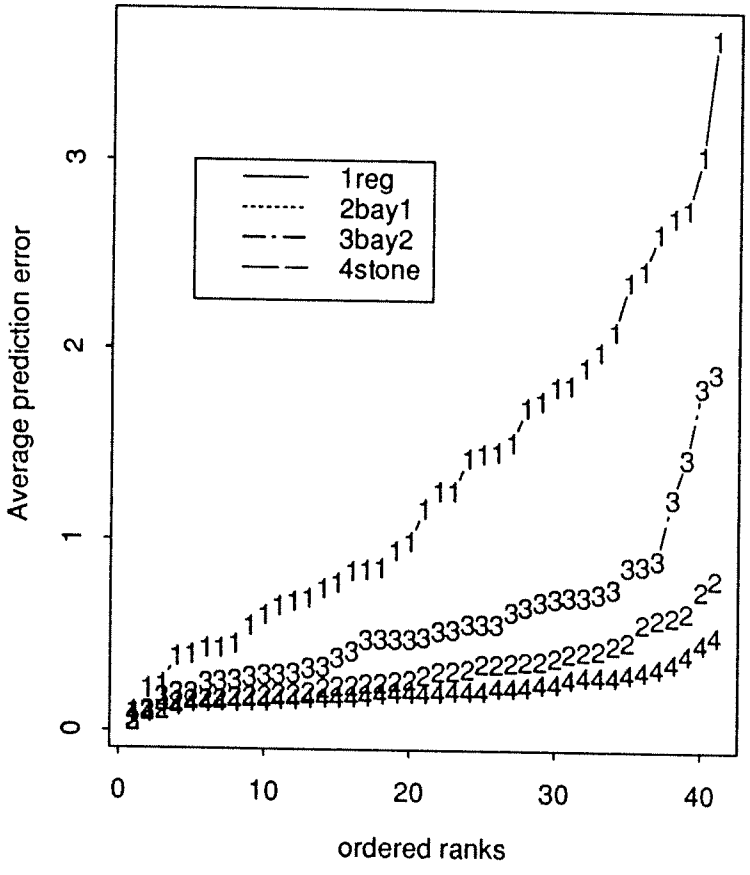
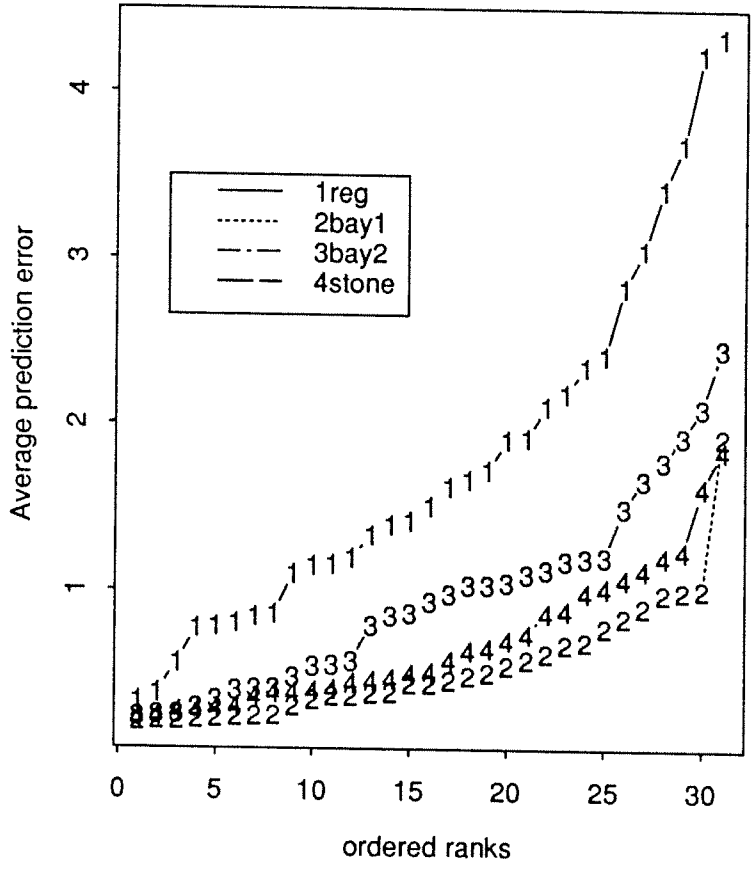


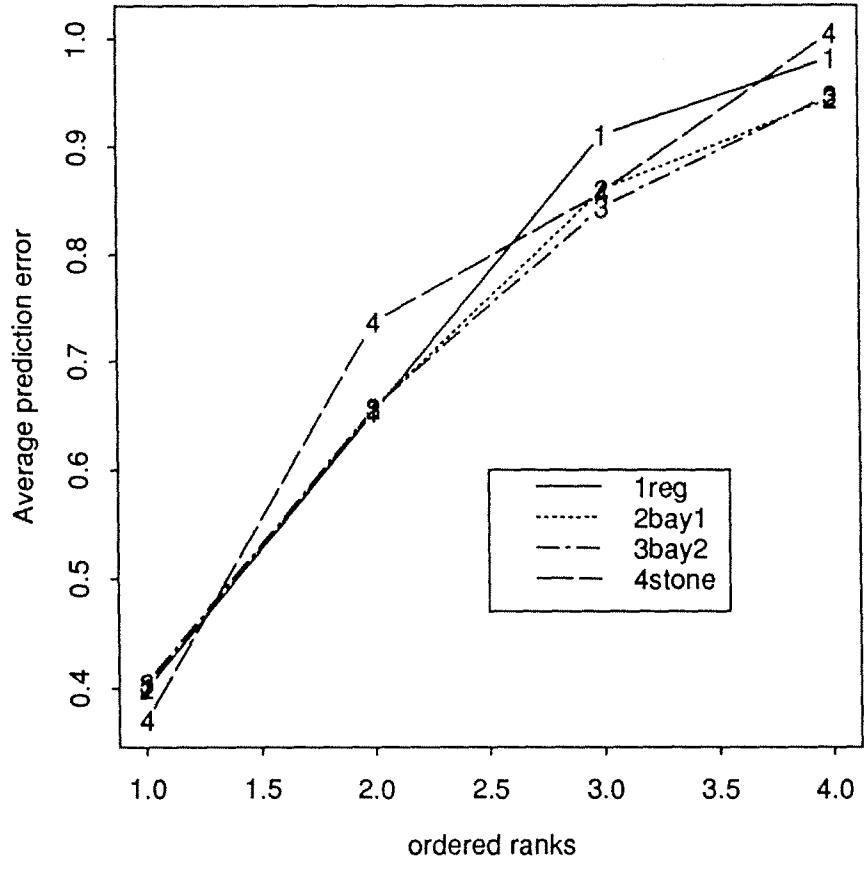
Figure A4.3(b): Cluster 2



Legend:
reg=regression, bay1= Bayesian alternative (1), bay2= Bayesian alternative (2)
Stone = Stone's procedure

Average Prediction Errors for Nitrate in Ascending Order

Figure A4.3(c): Cluster 3



Legend:
reg=regression, bay1= Bayesian alternative (1), bay2= Bayesian alternative (2)
Stone = Stone's procedure

BIOGRAPHICAL INFORMATION

NAME: KOMUNGOMA, S. K

MAILING ADDRESS: DEPARTMENT OF STATISTICS
UNIVERSITY OF DAR-ES-SALAAM
P.O BOX 35047
DAR-ES-SALAAM
TANZANIA

PLACE AND DATE OF BIRTH: BUKOBA, TANZANIA

EDUCATION (Colleges and Universities attended, dates, and degrees):

UNIVERSITY OF DAR-ES-SALAAM, 1980, BSC with Education

UNIVERSITY OF BRITISH COLUMBIA 1992 MSC.

POSITIONS HELD:

ASSITANT LECTURER

PUBLICATIONS (if necessary, use a second sheet):

AWARDS:

Complete one biographical form for each copy of a thesis presented to the Special Collections Division, University Library.