RRB-87-19

TIME SERIES ANALYSIS OF WATER CHEMISTRY IN CANADA AND IN NORWAY

by
E. Damsleth[1], D.C.L. Lam[2],
A.H. El-Shaarawi[2] and R.F. Wright[3]

[1]Norwegian Computing Centre
Box 335, Blindern, 0314
Oslo 3, Norway

[2]National Water Research Institute
Canada Centre for Inland Waters
Burlington, Ontario, L7R 4A6

[3]Norwegian Institute for Water Research
Box 333, Blindern, 0314
Oslo 3, Norway

May 1987
NWRI Contribution #87-109

ABSTRACT

This report summarizes the results of applying time series methods for developing stochastic models for various chemical data which were collected from Turkey Lakes Watershed in Ontario, Canada and the experimental watersheds used in the RAW (Reversing Acidification in Norway) project in Norway. Univariate and transfer function models were fitted to the pH, SO_4^{--} and M^{2+} data from Turkey Lakes Stations 1 and 2 and the Norwegian stations EGIL, KIM and ROLF. In addition, HCO_3^- data from the two Turkey Lake stations were also modelled. The results show that the univariate models explain about 80% of the total variability. The runoff contributes significantly as an explanatory variable for all the Turkey Lakes time series data and only for the pH at station KIM and SO_4^- and M^{2+} at station EGIL.

RÉSUMÉ

Le présent rapport résume les résultats de l'application de méthodes des séries chronologiques à l'élaboration de modèles stochastiques pour diverses données chimiques recueillies dans le bassin hydrographiques des lacs Turkey en Ontario, Canada et dans les bassins hydrographiques expérimentaux utilisés dans le projet norvégien RAIN (Lutte contre l'acidification en Norvège). Des modèles à variable unique et à fonction de transfert ont été adaptés aux données sur le pH, le SO_4^{--} et les M^{2+} provenant des stations 1 et 2 des lacs Turkey et des stations norvégiennes EGIL, KIM et ROLF. De plus, les données sur le HCO_3^- provenant des stations du lac Turkey ont également été modélisées. Les résultats indiquent que les modèles à variable unique expliquent environ 80% de la variabilité totale. Le ruissellement contribue sensiblement comme variable explicatoire pour toutes les données des séries chronologiques des lacs Turkey, seulement pour le pH à la station KIM, et pour SO_4^- et les M^{2+} à la station EGIL.

MANAGEMENT PERSPECTIVE

The acid rain problem is of significant concern in Europe and in North America. Substantial government support has led to massive research with this field. The idea of cost-sharing and effective information exchange among research program in different countries is foremost in the minds of scientists and policymakers. This paper presents the scientific results of two experiments conducted under such a concept. The first one is the RAIN (Reversing Acidification In Norway) Project which is financially supported by governments of Norway, Sweden, Canada and U.K. and is conducted at the southern part of Norway. The acid deposition is artificially controlled (reduced and also increased) and the chemistry of the manipulated watersheds is measured. Since the cost and the technology required are beyond the budget of any one of the participating countries, the concept of sharing and collaboration evolved. The second experiment is the Turkey Lakes Watershed study in Canada, in which several federal agencies and departments have participated, including Environment Canada, Canadian Forestry Service and Fisheries and Oceans. Again, sharing equipment and facilities is primary in the agenda.

This paper presents a new statistical method to examine the relationship of hydrology to the chemical responses in these watersheds. The results in Canada are compared to those in Norway, i.e. results for natural acidification are compared to those for artificial acidification. The statistical analyses are important for evaluating the sampling strategies used in these experiments.

PERSPECTIVE - GESTION

Le problème des pluies acides constitue une importante préoccupation en Europe et en Amérique du Nord. Un appui considérable de la part des gouvernements a entraîné des recherches massives dans ce domaine. L'idée de partage des coûts et d'échange d'information entre les divers programmes de recherche de pays différents est au premier plan chez les scientifiques et les décisionnaires. Le présent document présente les résultats scientifiques de deux expériences faites d'après ce concept. La première est le projet RAIN (Lutte contre l'acidification en Norvège) qui est appuyé financièrement par les governements de la Norvège, de la Suède, du Canada et du R.-U., et qui se déroule dans la partie méridionale de la Norvège. Les dépôts acides sont contrôlés artificiellement (réduits mais également augmentés) et la chimie des bassins hydrographiques manipulés est mesurée. Étant données que le coût et la technologie nécessaires dépassent les possibilités budgétaires et techniques de l'un ou l'autre des pays participants, le concept du partage et de la collaboration a donc progressé. La deuxième expérience est l'étude du bassin hydrographique des lacs Turkey au Canada à laquelle plusieurs organismes et ministères fédéraux ont participé, notamment Environnement Canada, le Service canadien des forêts et Pêches et Océans. Ici aussi, il est primordial de partager l'équipement et les installations.

Ce document présente une nouvelle méthode statistique pour étudier le rapport entre l'hydrologie et les réactions chimiques de ces bassins hydrographiques. Les résultats obtenus au Canada sont comparés avec ceux de la Norvège, c.-à-d. que les résultats de l'acidification naturelle sont comparés à ceux de l'acidification artificielle. Les analyses statistiques sont importantes pour évaluer les stratégies d'échantillonage utilisées dans le cadre de ces expériences.

1. INTRODUCTION

During the last 10-15 years the environmental problems related to the effects of acid rain have become more and more evident in most industrialized areas around the world. This, combined with an increaased public awareness of the problems and substantial governmental support has led to massive research within this field, worldwide.

Watershed acidification is a complex process, due to the many interacting chemical reactions and processes involved. Several scientists have spent considerable effort on developing physically-based, deterministic hydrochemical models to incoroporate these reactions and processes into large acidification models (Chen et al. 1982, Christophersen et al. 1982, Cosby et al. 1984, Lam and Bobba 1984). Parallel to this work, which has been performed mostly by chemists and/or physicists, statisticians have spent a lot of effort on stochastic models of watershed hydrology and chemistry. Hipel and McLeod (1987) give an excellent summary of the state of the art in this field. Whitehead et al. (1984) present one attempt to combine the two schools, and Damsleth (1986) provides an application to a set of Norwegian data.

The scope of the present study was to develop stochastic time series models, on a daily basis, for various chemical time series observed in the Turkey Lakes watershed in Ontario, Canada and in the experimental watersheds used in the RAIN (Reversing Acidification in Norway) project in Norway. The purpose of the study was to see how well these models fitted the observed data, how similar (or different) the models were between watersheds and countries as well as from one chemical variable to another.

The outline of the report is as follows: Section 2 gives a description of the watersheds involved in the study and the various chemical variables subject to analysis. Section 3 presents a brief summary of time-series analysis theory, without going into any details. Section 4 describes the sampling frequency of the various series in the study, which varies a lot. The main results from the analysis are presented in Section 5, while Section 6 gives a summary of our findings. The report is ended with an appendix which extends the application of the Kalman filter technique to non-stationary ARIMA-models with missing observations. This extension may be of interest in other applications as well.

2. THE WATERSHEDS AND THEIR CHEMISTRY

The Turkey Lakes Wateshed is an undisturbed, forested basin located on the Precambrian Shield approximately 50 km north of Sault Ste. Marie, Ontario, Canada. The watershed contains a series of five lakes over a total area of 10.5 km^2. A series of sampling stations (TURKEY1-TURKEY5) are located along the main stream. The lowest elevation is 245 m and the highest, 645 m. The soil is composed of loam and sandy loam at higher elevations (depth = 0.2 to 1 m) and of gravelled till and fine-grained till at lower locations (depth = 1 to 10 m). Atmospheric deposition of acid is moderate in this region, with an average pH of 4.5 and an annual sulphate load of about 20 kg SO_4 ha^{-1}. Ionic budget (Jeffries et al. 1986) showed that atmospheric deposition directly to the lakes' surfaces was the principal input pathway for H^+ and NH_4^+, whereas $SO_4^=$ and NO_3^- were derived mainly from the surrounding terrestrial basin and upstream lakes. Strong spatial gradients of increasing Ca^{++}, Mg^{++} and alkalinity were observed in the downstream direction, i.e. from TURKEY1 to TURKEY5. These gradients were related to the increase in groundwater flow to the stream at lower locations (Lam et al. 1986). Thus, the mean pH for TURKEY1 and TURKEY2 are 6.14 and 6.48, as compared to values between 6.8 and 7.2 in the lower stations. To simplify the comparison, however, we use the data from TURKEY1 and TURKEY2 only in the time series analysis. In general, the water chemistry is strongly affected by the watershed hydrology and meteorological episodes. Of the four years (1981-1984) of time series data, 1981 had a snowmelt episode in April followed by several heavy rains in June and a long dry summer and autumn. In 1982, snowmelt occurred late in May, followed by a dry summer and a very wet autumn. In 1983, a warm winter led to several melting and thawing sequences, a dry summer and a wet autumn. The weather of 1984 was quite similar to 1982.

The RAIN experiment comprises two parallel manipulations in which acid deposition is experimentally changed at whole catchments (Wright et al. 1986). At Sogndal in western Norway, a pristine headwater catchment is artificially acidified. At Risdalsheia in southernmost Norway, ambient acid precipitation is excluded by a roof at the KIM catchment (area=860 m^2). The precipitation is collected and cleaned by ion-exchange and applied beneath the roof. Natural levels of sewater salts are re-added to the applied water. Two nearby catchments, EGIL (area = 400 m^2, also with a roof but the applied precipitation is not cleaned) and ROLF (area = 200 m^2, with no roof), serve as control. At this stage, we concentrate on the time series data of Risdalsheia, as those of Sogndal need further compilation.

The Risdalsheia catchments are sparsely forested and characterized by thin (average depth = 15 cm) and patchy podzodic soils on siliceus gneissic-granitic bedrock. The rain pH is 4.2 and the stream pH is about 4.0 to 4.4, accompanied by elevated concentrations of labile inorganic aluminium. Acid exclusion began in June 1984, with a dry summer followed by a wet autumn, quite typical of the area. By November 1985, a total of 1100 mm of precipitation had been cleaned at the KIM catchment. In fact, by December 1984 at the onset of winter sulphate concentrations at KIM were about 80 µeq/L as compared to 100 to 120 µeq/L at EGIL and ROLF, while the pH levels were about 0.2 units higher. The Ca^{++} + Mg^{++} (or M^{2+}) concentration from the three catchments is generally about one fifth of those at the Turkey Lakes headwater catchements.

Thus, while the Turkey Lakes Watershed is subjected to moderate acid load and protected by relatively thicker soil, the Risdalsheia

Watershed receives stronger acid input with less soil buffering. The weather pattern and hydrology in both cases, however, influence the water chemistry significantly. Since both watersheds have three to four years of data, a statistical investigation of the relationships between hydrology and chemistry is possible. To simplify the comparison, we examine the four chemical variables, pH, $SO_4^=$, M^{2+} and HCO_3^- only.

3. TIME SERIES MODELS

Suppose that we have observed a time series at equidistant points in time, so that y_t denotes the observed value at time t. As an example, y_t may be daily measurements of a chemical variable, say pH, in a river. It is common in environmental series not to have equidistant sampling intervals. This generates additional problems in the analysis, which will be considered later on. For the time being, assume that the series is completely observed.

3.1 Univariate Models

The first step in a time series analysis is to build a univariate model for the series. The purpose of these models is to separate the series in two components: one which can be predicted from the series own past, and one unpredictable, random component. Thus, we seek a representation as

3. TIME SERIES MODELS

Suppose that we have observed a time series at equidistant points in time, so that y_t denotes the observed value at time t. As an example, y_t may be daily measurements of a chemical variable, say pH, in a river. It is common in environmental series not to have equidistant sampling intervals. This generates additional problems in the analysis, which will be considered later on. For the time being, assume that the series is completely observed.

3.1 Univariate Models

The first step in a time series analysis is to build a univariate model for the series. The purpose of these models is to separate the series in two components: one which can be predicted from the series own past, and one unpredictable, random component. Thus, we seek a representation as

$$y_t = f\,(y_{t-1}, y_{t-2}, \ldots ; a_{t-1}, a_{t-2}, \ldots) + a_t$$

where a_t is the random, unpredictable part of y_t while f denotes the part which can be predicted from the past values of y and a. The sequence $\{a_t\}$ is supposed to be white noise, that is a series of independent, identically distributed random (normal) variables with mean zero.

Within the class of ARMA (Auto Regressive Moving Average) models, the function f is assumed to be linear in the y's and a's, so that an ARMA(p,q) model is given by

$$y_t = \mu + \phi_1 (y_{t-1} - \mu) + \phi_2 (y_{t-2} - \mu) + \dots + \phi_p (y_{t-p} - \mu)$$
$$- \theta_1 a_{t-1} - \dots - \theta_q a_{t-q} + a_t \qquad (3.1)$$

Here μ represents the mean of the series, $\phi_1,\dots,\phi_p$ are the autoregressive (AR) parameters and $\theta_1,\dots,\theta_q$ are the moving average (MA) parameters. The model (3.1) can be rewritten as

$$(1 - \phi_1 B - \dots - \phi_p B^p)(y_t - \mu) = (1 - \theta_1 B - \dots - \theta_q B^q)\, a_t \qquad (3.2)$$

where B is the backwards shift operator which operates on the time index so that $B^k y_t = y_{t-k}$.

In this model, the y_t series is assumed to be stationary, that is: the statistical properties of the series do not change with time. More specific, the series is assumed to vary around a constant mean, with a constant variance. Series with changes in levels, local or global trends, etc. are frequently encountered, resulting in a non-stationary series. Since most of the theory behind time series analysis is developed for stationary series, this may constitute a problem. However, many non-stationary series can be transformed to a stationary series by differencing.

Thus, if y_t is the observed series, we look at the series given by $y_t-y_{t-1}=(1-B)y_t$ or $(y_t-y_{t-1})-(y_{t-1}-y_{t-2})=(1-B)^2y_t$. The general ARIMA model of orders (p,d,q) can then be written

$$(1 - \phi_1 B - \dots - \phi_p B^p)(1 - B)^d \dot{y}_t = (1 - \theta_1 B - \dots - \theta_q B^q) a_t \quad (3.3)$$

p, d and q are the order of the AR-term, the number of differences required to obtain stationarity and the order of the MA-term, respectively. $\dot{y}_t$ denotes the actually observed value of y_t if $d>0$, if $d=0$ $\dot{y}_t = y_t - \mu$, the deviation from the mean level. The outline of a univariate analysis is as follows:

1. Identify a tentative model using plots of series, and the autocorrelation and partial autocorrelation functions.
2. Estimate the parameters in the model using maximum likelihood techniques, and calculate the resultant noise series (the residuals).
3. Perform some diagnostic checks on the residuals, to make sure that they are white noise. If they are, then the model is okay, otherwise the diagnostic checking will suggest an improved model, with which to return to step 2.

It is beyond the scope of this report to describe the details in the model building and estimation procedures which can be found in Box and Jenkins (1974).

3.2 Transfer Function Models

In many situations the time series under study is known or assumed to be influenced by one or several other variables. There is, for example, evidence in the literature that the runoff affects the pH (Damsleth 1986). When this relationship is one sided, so that there is no feed-back in the system, such effects may be modelled within the framework of single-input or multiple-input transfer function models. The runoff to pH relationship is a clear-cut case; it is hard to argue that the pH affects the runoff!

A single-input transfer function model can be written

$$\begin{aligned} y_t &= \nu_0 x_t + \nu_1 x_{t-1} + \ldots + n_t \\ &= (\nu_0 + \nu_1 B + \nu_2 B^2 + \ldots) x_t + n_t \qquad (3.4) \end{aligned}$$

where y_t is the output series (e.g. pH), x_t is the input series (e.g. runoff) and n_t is a noise term which is assumed to follow an ARIMA model as in (3.3).

The weights $\nu_0, \nu_1, \ldots$ in (3.4) are the impulse response weights of the model, and describe how a change in the input is transferred to the output. Some of the first ν-values may be 0, implying a delayed response from input to output. Usually the ν_k-values will tend to die out as k increases, and be effectively zero from a certain k on. Sometimes a fairly large number of ν's is required to give an adequate representation of the transfer function. If this is the case, it may frequently prove useful to reduce the number of parameters by writing

$$\nu_0 + \nu_1 B + \dots = \frac{\omega_0 - \omega_1 B - \dots - \omega_r B^r}{1 - \delta_1 B - \dots - \delta_s B^s} \qquad (3.5)$$

where the high (or infinite) order polynomial $\nu_0 + \nu_1 B + \dots$ is replaced by the ratio of two low order polynomials, which may lead to a considerable reduction in the number of parameters. A complete single input transfer function model may then be written as

$$y_t = \frac{\omega(B)}{\delta(B)} x_t + \frac{\theta(B)}{\phi(B)(1 - B)^d} a_t \quad , \qquad (3.6)$$

where $\omega(B)$, $\delta(B)$, $\theta(B)$ and $\phi(B)$ are polynomials in the backwards shift operator B, of order r, s, q and p, respectively.

The generalization of (3.6) to the case with several input series is straightforward. In that case

$$y_t = \sum_{i=1}^{m} \frac{\omega_i(B)}{\delta_i(B)} x_{it} + \frac{\theta(B)}{\phi(B)(1 - B)^d} a_t \qquad (3.7)$$

where m is the number of input series, and $\omega_i(B)/\delta_i(B)$ is the transfer function for the ith input series $\{x_{it}\}$, i=1, ..., m.

As in the unvariate case, there are certain techniques (Box and Jenkins 1974) to identify, estimate and diagnose transfer function models, where the cross-correlation function between the output and the input(s) plays an important role.

3.3 Missing Observations

Environmental data are only rarely observed at equidistant time intervals. In the present applications, the sampling interval varies from 1 to more than 30 days. This makes the ordinary techniques, as given in Box & Jenkins (1974), inapplicable. Difficulties occur in the univariate as well as the transfer function situation, both when it comes to identifying the structure of the model and in the estimation stage. Our approach has been to regard the data as daily time series with a large number (80%) of missing data. In this way we utilize all the information, without throwing any data out.

3.3.1 Univariate Model Building

The identification of the univariate model was performed in the usual way, where the autocorrelation functions were computed as in Barham & Dunstan (1982). Then the parameters of the identified model are estimated by the technique of Jones (1980) which is outlined in Appendix A. The technique reformulates the ARMA model in a Kalman filter framework, and utilizes the ability of the Kalman filter to handle missing observations in an automatic way. In the appendix Jones' technique is generalized from the ARMA to the ARIMA-models.

As a result of the missing observations, the residuals are not identically distributed and the ordinary diagnostic tests are difficult to perform. Instead the method of overfitting was used in a stepwise manner to determine an adequate model for fitting the data, where the fit was measured by the Akaike information criterion (AIC), see e.g. Shibata (1985) or Akaike (1972, 1973, 1974). The same criterion was used to differentiate between stationary (without differencing) and non-stationary (with differencing) models.

3.3.2 Transfer Function Models

It is difficult to fit a general transfer function model in the presence of lots of missing data. The technique consists of two steps: (1) a (non-white) noise series $\{n_t\}$ is calculated from

$$n_t = y_t - \frac{\omega(B)}{\delta(B)} x_t \qquad (3.8)$$

and, (2) a univariate model is fitted to the $\{n_t\}$ series. The $\delta(B)$ and $\omega(B)$ generate problems when there are lots of missing values in the output and the input series, respectively.

In the present study, fortunately, the input series (mostly runoff) contains only a few missing observations, while the output series includes a massive amount of missing data. We therefore restricted ourselves to transfer fucntions with $\delta(B) \equiv 1$, so that there is no denominator polynomial in the transfer function. From (3.8) n_t can be computed by

$$n_t = y_t - \omega_0 x_t - \omega_1 x_{t-1} - \dots - \omega_r x_{t-r} \qquad (3.9)$$

when y_t, x_t, x_{t-1}, ..., and x_{t-r} are observed. Otherwise n_t is considered to be missing in the subsequent univariate analysis.

The order of the transfer function is determined by increasing r in (3.9) until no improvement of fit has occurred, taking some knowledge of the physical processes involved into account.

4 SAMPLING FREQUENCY

In the present study, runoff is measured almost daily, with much fewer missing observations than the chemical measurements. Generally, more observations are taken during "periods of interest", where "periods of interest" are not well defined, but tends to include the snowmelt period.

4.1 Sampling at Turkey Lakes

Turkey Lakes data were collected from 1981 to 1984. Runoff measurements started on February 26, 1981, for Station 1. There are 48 runoff observations missing at Station 1, all during July, August and September, while all runoff data are available for Station 2.

The observation pattern of the chemistry, given as number of observations per month, is shown in Figure 4.1 for Stations 1 and 2. The chemical variables are normally sampled simultaneously, and the figures show the pattern for all four chemical variables. Except for pH, the unit used for the concentration is μ mole/L. On the few months where the number of observations differed between the variables, the mean is plotted in the figures. The sampling patterns are very similar for the two stations, and show a significant increase in the intensity from 1981 to 1982, and a clear seasonal pattern with much more frequent sampling during the snowmelt period.

Except for the intense study periods during the spring, the basic sampling interval is weekly. There are, however, large variations in the time between successive observations, which can be seen from the histograms in Figure 4.2, which show a peak at 1 day and a not very distinct bump around 7 days. The average sampling interval was 5.5 days for Turkey 1 and 5.6 days for Turkey 2.

4.2 Sampling in the RAIN Project

The observations started on 840318, 840320 and 840415 for EGIL, KIM and ROLF, respectively. For ROLF, however, runoff data were not available before 841101. The last observations were taken on 861220 for all three stations. The runoff was observed daily, and there were five missing observations in each runoff-series, not including the ones missing for the first 6-1/2 months at ROLF.

There is a small problem with the runoff measurements in the RAIN project. The runoff is measured in units of one tank, where one tank represents 0.930 mm/day at EGIL, 0.713 mm/day at KIM and 1.62 mm/day at ROLF. The observations are converted to mm/day in this analysis, but the resulting observations will thus no longer be measured at a continuous scale. Further, during dry periods, the runoff will frequently not exceed one tank, resulting in zero runoff, which is not consistent with the fact that the chemistry in the runoff has been analysed on such days. We do not expect this to constitute any major problem, but it may, to a certain extent, weaken the potential relationship between the runoff and the chemical variables. To facilitate the comparison, the runoff and chemical variables are expressed in the same units as those of Turkey Lakes.

The sampling patterns for the three RAIN stations are shown in Figure 4.3, and they are even more erratic than those for Turkey Lakes. In March 1984, there were 29 and 30 observation days at EGIL and ROLF, while there are several months without observations at all three stations. In fact, the range is even larger, as there are several days with more than one observation. In some cases, the chemistry was measured hourly for almost 24 hours. However, analysing hourly data was beyond the scope of this work, so we picked one observation at random in the cases where several observations were made on the same day. The pattern at ROLF differs somewhat from EGIL and KIM, as the high frequency sampling in 1984 did not occur.

The histograms of the time intervals between observations are shown in Figure 4.4, and confirm the conclusions above. There are peaks at 1 day for EGIL and KIM, and to a smaller extent for ROLF. Otherwise the distributions of sampling intervals are fairly flat, showing great variations in the assumed weekly sampling scheme. There is also a larger proportion of intervals >15 days when the RAIN-data are comapred to Turkey Lakes. This is due to the long non-sampling periods during mid-winter in the RAIN project. The average sampling interval is 5.4, 5.2 and 8.8 days for EGIL, KIM and ROLF, respectively.

5. RESULTS FROM THE ANALYSIS

Univariate time series and transfer function models were fitted for the chemical variables pH, SO_4^{--} and M^{2+} using the data from Turkey Lake Stations 1 and 2, and the Norwegian stations EGIL, KIM and ROLF. We have also analysed the HCO_3^- from the two Turkey Lake stations. This chapter summarizes the findings.

5.1 Main Results

- Stationary models fit the series better than non-stationary models for all series, but the difference is not substantial.
- A univariate ARMA(1,1)-model with AR-parameter fairly close to 1 fits most of the series under study satisfactorily, though we found it necessary to include MA-terms of order 2 and 3 in some cases. The univariate model "explains" about 80% of the variance in most of the series.
- The runoff contributes significantly as an explanatory variable for all the Turkey Lakes series. For the RAIN series the runoff played a significant role only for the pH at station KIM and the SO_4^- and M^{2+} at Station EGIL. In all cases the contribution is

small, increasing the percent of variation explained by 1-3%.

- The variability, measured as a coefficient of variation, is much larger in the Norwegian data, as far as SO_4^{--} and M^{2+} are concerned. The pH level is much lower in the Norwegian data, and thus also the coefficient of variation.
- For the Turkey Lake series, rainfall and snowmelt does not contribute any additional information when the runoff is known. For the RAIN data these data were not available.

5.2 Models for the Runoff

The runoff consitutes the explanatory variable for all the chemical series under study. Table 5.1 gives some basic statistics for the runoff at the five stations, and the univariate models are summarized in Table 5.2. The models are all ARMA(1,2), given by

$$(1 - \phi_1 B)(x_t - \mu) = (1 - \theta_1 B - \theta_2 B^2)\, a_t \tag{5.1}$$

or simplications thereof. In (5.1), x_t denotes the observed runoff on day t, μ is the mean runoff and $\{a_t\}$ denotes a sequence of independent, identical normally distributed random variables, with variance σ_a^2, i.e. a white noise series. The operator B is the backwards shift operator, so that $B^k x_t = x_{t-k}$. ϕ_1 is the autoregressive parameter and θ_1 and θ_2 are the moving average parameters. In Table 5.2, R^2 denotes the proportion of the variance "explained" by the model, $R^2 = 1-\sigma_a^2/\sigma_x^2$, where σ_x^2 is the variance of the observed series.

Table 5.1 Basic statistics for the runoff (mm/d) at the five stations

Station	No. of Observations	Mean	Variance	Coefficient of Variation
Turkey 1	1357	.252	.249	1.98
Turkey 2	1461	.206	.127	1.73
EGIL	996	.189	.169	2.18
KIM	996	.253	.276	2.08
ROLF	775	.345	.676	2.38

Table 5.2 Parameters in the univariate model (5.1) for the runoff - means that the parameter was not significantly different from 0 and was excluded from the estimation

Station	μ	ϕ_1	θ_1	θ_2	σ^2_a	R^2
Turkey 1	.246	.910	.38	.18	.136	.45
Turkey 2	.205	.891	-.07	.20	.031	.76
EGIL	.189	.955	.64	-	.134	.21
KIM	.250	.858	.50	-	.184	.33
ROLF	.344	.849	.53	.14	.344	.49

There are several interesting features in Tables 5.1 and 5.2:

- The variance for Turkey 2 is less than that for Turkey 1, and this becomes even more pronounced when the residual variances, σ_a^2, are compared. This implies that the runoff at Turkey 2 is much more predictable than at Turkey 1. This makes sense considering that Turkey 1 is a headwater station and thus subjected to direct influence of the variation of the precipitation episodes.
- The three RAIN runoff series differ substantially, both with respect to mean, variance and residual variance. This is probably due to the differences in watershed areas.
- The univariate models give, in general, a better fit for the Turkey Lakes series than for the RAIN series. This is mostly due to the size of the watersheds, but some of the reason may be found in the better precision in the Turkey Lakes measurements.

5.3 Models for pH

Some basic statistics on the pH-measurements at the five stations are given in Table 5.3.

Table 5.3 Basic statistics for pH at the five stations

Station	No. of Observations	Mean	Variance	Coefficient of Variation
Turkey 1	268	6.14	.063	.04
Turkey 2	270	6.48	.054	.04
EGIL	187	4.08	.028	.04
KIM	193	4.11	.015	.03
ROLF	112	4.01	.024	.04

5.3.1 Univariate Models

All the pH series are well fitted by univariate ARMA-models. An ARMA(1,1) fits well for all the stations except EGIL, where an ARMA(1,3) was necessary. Thus, all the series can be modelled by

$$(1 - \phi_1 B)(y_t - \mu) = (1 - \theta_1 B - \theta_2 B^2 - \theta_3 B^3)\, a_t \qquad (5.2)$$

or simplifications thereof. In (5.2) y_t denotes the observed pH on day t. Table 5.4 shows the estimated parameter values for the five stations.

The most striking features in Tables 5.3 and 5.4 are the large difference in the means between the Canadian and the Norwegian stations, the almost identical models for the two Turkey Lakes stations, and the peculiar, large value of θ_3 for EGIL. It is also of interest to notice that the AR parameter in general is lower for the

Table 5.4 Parameters in the univariate model (5.2) for pH - means that the parameter was not significantly different from 0 and was excluded from the estimation

Station	μ	ϕ_1	θ_1	θ_2	θ_3	σ_a^2	R^2
Turkey 1	6.19	.971	.50	-	-	.014	.78
Turkey 2	6.48	.970	.54	-	-	.012	.79
EGIL	4.07	.963	-.25	.38	.60	.0082	.71
KIM	4.11	.934	.35	-	-	.0032	.79
ROLF	4.01	.920	-	-	-	.0040	.83

RAIN series, so the "memory" of the Norwegian series is shorter. This can also be explained by the size and nature of the watersheds.

5.3.2 Transfer Function Model

Hydrological theory predicts that the runoff should affect the pH. There may be a diluting effect, so that increased runoff leads to an increase in pH, or increased runoff may lead to a decrease in the pH if the runoff is due to a very acid rainfall or melting of snow with low pH.

For all the series we fitted a transfer function model of the form

$$y_t = \alpha + (\nu_0 + \nu_1 B + \nu_2 B^2 + \ldots + \nu_r B^r)\, x_t + n_t \qquad (5.3)$$

In (5.3), y_t is the observed pH on day t and x_t the observed runoff. α is an intercept constant, and the parameters ν_0, ν_1, ... are the impulse respond weights, which describe how a change in runoff at day t affects the pH at the same day, the day after, two days after, etc. In (5.3), n_t denotes a noise term, which (as opposed to a traditional regression model) is not white, but follows an ARMA-model similar to the one given in (5.2). Table 5.5 gives a summary of the transfer function parameters for the five stations. The noise models were only marginally different from those given in Table 5.4, and are not shown.

Table 5.5 Transfer function parameters in model (5.3) between runoff and pH. - denotes a parameter which was not significantly different from 0, and thus was excluded from the estimation. A parameter marked with a * is less than two times its standard deviation, and thus not significant at a level about 5%.

Station	α	ν_0	ν_1	ν_2	ν_3	σ^2_a	R^2
Turkey 1	6.23	-.130	-	-	-	.011	.83
Turkey 2	6.52	-.057	.007*	-.031*	-.096	.011	.79
EGIL	4.07	-	-	-	-	.0082	.71
KIM	4.10	-.014	-.004*	.017	-	.0029	.81
ROLF	3.99	-	.015	-	-	.0037	.85

The most interesting feature of Table 5.5 is its lack of consistency. The only station where the runoff gives a noticeable contribution is Turkey 1, where the introduction of the runoff increases the R^2 from .78 to .83. Otherwise the improvement is just marginal, and the structure of the models as well as the sign of the parameters differs wildly, and does not lend themselves to any easy interpretation. It is worth noticing that for Turkey 1, where the impact of the runoff was the most significant, the effect is instantaneous and negative. Thus, an increase in the runoff leads to a drop in pH, in favour of the snowmelt concept.

From Turkey Lakes, we have access to data on the rainfall and snowmelt as well. When these series were used as the only input to the transfer function model, we found significant effects for most of the series. However, the significance disappeared completely when the rainfall and snowmelt was introduced in addition to the runoff, showing that the rainfall and snowmelt data contain no additional information, considered within a transfer function framework.

5.3 Models for SO_4^{--}

Summary statistics for the SO_4^{--} measurements are presented in Table 5.6.

The level of SO_4^{--} is somewhat lower in the Norwegian watersheds, but the difference in variances and coefficient of variation is striking. As expected, since the precipitation is cleaned at KIM, a lower SO_4^{--} concentration is expected. The result also confirmed that SO_4^{--} is adsorbed more in Risdalsheia than in Turkey Lakes.

Table 5.6 Summary statistics for SO_4^{--} (μ mole/L)

Station	No. of Observations	Mean	Variance	Coefficient of Variation
Turkey 1	251	62.9	68.8	.13
Turkey 2	254	62.5	65.7	.13
EGIL	187	58.7	1094.	.56
KIM	193	45.7	658.	.56
ROLF	112	54.4	1021.	.59

5.3.1 Univariate Models

The five SO_4^{--} series can all be modelled within the framework of an ARMA(1,1) model, given by (5.1), where y_t now denotes the observed SO_4^{--} concentration on day t. The estimates parameters are given in Table 5.7.

Table 5.7 Estimated parameters in the univariate model (5.1) for SO_4^{--} at the five stations. - denotes a parameter which was not significantly different from 0, and was excluded from the estimation

Station	μ	ϕ_1	θ_1	σ^2_a	R^2
Turkey 1	62.9	.971	.37	8.17	.88
Turkey 2	62.4	.971	.23	4.86	.93
EGIL	69.9	.932	-.34	115.	.90
KIM	43.3	.947	.30	105.	.84
ROLF	59.5	.899	-	212.	.79

The models for the two Turkey Lakes stations are very similar for SO_4^{--}, as they were for pH. The residual variance is smaller for Turkey 2, and R^2 is larger, so that the SO_4^{--} in Turkey 2 is more predictable compared to Turkey 1. There is less agreement between the models for the three RAIN series. The AR-parameters are fairly similar, but the MA parameters differ wildly. The differences in residual variance and R^2 are also substantial, though all three series have fairly high R^2 values.

5.3.2 Transfer Function Models

We also fitted transfer function models to the five SO_4^{--} series, using runoff as input. The model was of the form (5.3), where y_t now denotes the observed SO_4^{--} concentration, and the results are presented in Table 5.8.

Table 5.8 Transfer function parameters in model (5.3) between runoff and SO_4^{--}. - denotes a parameter which was not significantly different from 0, and thus was excluded from the estimation. A parameter marked with a * is less than two times its standard deviation, and thus not significant at about 5% level.

Station	α	ν_0	ν_1	ν_2	ν_3	σ^2_a	R^2
Turkey 1	64.2	-1.50	.01*	-.65	-	6.43	.91
Turkey 2	63.0	-1.68	-1.07	-	-	4.31	.93
EGIL	75.9	-4.12	-5.42	-6.02	-6.72	105.	.90
KIM	43.3	-	-	-	-	105.	.84
ROLF	59.5	-	-	-	-	212.	.79

The univariate models for the noise are again almost identical to those given in Table 5.7, and are not shown.

From the knowledge of the physical and chemical processes involved, one would expect the relationship between runoff and SO_4^{--}, if any, to be instataneous and negative. It is thus surprising to find substantial delayed effects as well, especially for EGIL. It is also worth noticing that we found no significant effect of the runoff on the SO_4^{--} for KIM and ROLF.

5.4 Models for M^{2+}

The basic statistic on the M^{2+} measurements at the five stations are given in Table 5.9.

Table 5.9 Summary statistics for M^{2+} (μ mole/L)

Station	No. of Observations	Mean	Variance	Coefficient of Variation
Turkey 1	256	94.7	150.	.13
Turkey 2	261	115.6	136.	.10
EGIL	185	22.9	221.	.65
KIM	189	19.7	118.	.55
ROLF	111	22.7	134.	.51

The Turkey Lakes and the RAIN series differ significantly, as can be seen from the table. The RAIN concentration is only about 1/5 of Turkey Lakes, while the variances are of the same magnitude, giving the RAIN series a much larger coefficient of variation. This result is attributable to the thicker soil layer and more $CaCO_3$ input from the soil to the streams at Turkey Lakes.

5.4.1 Univariate Models

All five series can be fitted by the ARMA(1,2) model (5.1) or simplifications thereof, where y_t now represents the observed M^{2+} concentration at day t. The estimated parameters are shown in Table 5.10.

Table 5.10 Estimated parameters in the univariate model (5.1) for M^{2+} at the five stations. - denotes a parameter which was not significantly different from 0, and was excluded from the estimation

Station	μ	ϕ_1	θ_1	θ_2	σ^2_a	R^2
Turkey 1	96.5	.965	.33	-	25.0	.83
Turkey 2	116.5	.953	.55	-	50.7	.63
EGIL	25.8	.888	-.61	-.37	17.8	.92
KIM	20.9	.956	-	-	12.2	.90
ROLF	24.4	.892	-	-	30.9	.77

In Table 5.10, note that the residual variance for Turkey 2 is twice that of Turkey 1, and that R^2 is accordingly less. Thus, while Turkey 2 is the smoothest and most predictable when pH or SO_4^{--} is concerned, the situation is the opposite with regards to M^{2+}. The RAIN series again differ substantially, both among themselves and from the Turkey Lakes series. By and large, however, the RAIN series are the more predictable.

5.4.2 Tranfer Function Models

As the M^{2+} is released from the soil, one would expect a certain delayed effect of runoff on M^{2+}. This is to a certain extent supported by Table 5.11, which shows the estimated parameter values in model (5.3) for the transfer function between runoff and M^{2+}.

Table 5.11 Transfer function parameters in model (5.3) between runoff and M^{2+} - denotes a non-significant parameter which has been excluded from the estimation.

Station	α	ν_0	ν_1	ν_2	ν_3	σ_a^2	R^2
Turkey 1	98.6	-2.98	-1.53	-1.58	-1.66	20.1	.87
Turkey 2	119.6	-5.49	-6.03	-	-	44.4	.67
EGIL	27.8	-2.59	-2.69	-2.18	-	16.4	.93
KIM	20.9	-	-	-	-	12.2	.90
ROLF	26.1	-	-	-	-	30.9	.77

As was the case for the pH and SO_4^{--}, the noise models changed only marginally from those given in Table 5.10.

The effect of the runoff on the M^{2+} is well pronounced in the two Turkey Lakes series, and the delayed effect is clear, particularly at Turkey 1. A small, but significant, improvment was found at EGIL as well, while we found no significant effect of the runoff on the M^{2+} at KIM or ROLF.

5.5 Models for HCO_3^-

We have not had access to HCO_3^- measurements from the RAIN-project, so the analysis in this case is limited to the two Turkey Lakes stations. Table 5.12 gives the basic statistics for the two HCO_3^- series.

Table 5.12 Summary statistics for HCO_3^- (μ mole/L)

Station	No. of Observations	Mean	Variance	Coefficient of Variation
Turkey 1	256	40.0	416.	.51
Turkey 2	258	88.3	838.	.33

The HCO_3^- concentration at TURKEY2 is higher than TURKEY1 and is part of the spatial gradient of increasing HCO_3^- from headwater to downstream lakes at Turkey Lakes.

5.5.1 Univariate Models

Both series can be well fitted by ARMA(1,1) models, given by

$$(1 - \phi_1 B)(y_t - \mu) = (1 - \theta_1 B)\, a_t \qquad (5.4)$$

where y_t is the observed HCO_3^- concentration on day t, and the rest of the equation is defined as before. Table 5.13 gives the estimated parameters.

Table 5.13 Univariate model parameters for HCO_3^-.

Station	μ	ϕ_1	θ_1	σ^2_a	R^2
Turkey 1	46.3	.966	.54	122.	.71
Turkey 2	96.8	.973	.59	229.	.73

The correlation parameters and the R^2 values are very similar for the two series, while the difference in level and variance remains.

5.5.3 Transfer Function Model

As for M^{2+}, we expect a certain delayed effect on the HCO_3^- from the runoff. The parameters in the transfer function model (5.3), which are shown in Table 5.14, confirms this.

Table 5.14 Parameters in the transfer function between runoff and HCO_3^-. The parameter marked * does not exceed two times its standard deviation, and is thus not significant about the 5% level.

Station	α	ν_0	ν_1	ν_2	ν_3	σ_a^2	R^2
Turkey 1	49.3	-5.12	-3.79	-1.64	-3.93	115.	.72
Turkey 2	103.5	-8.44	-10.3	.03*	-11.2	202.	.76

Except for the somewhat peculiar estimate of ν_2 for Turkey 2, the values agree with the a priori expectations, and the introduction of runoff to the model leads to a significant, though not substantial, improvement in the fit. The models for the noise remains practically unchanged from those given in Table 5.13.

6. CONCLUSIONS

The application of Kalman filter technique in the time series analysis of two different sets of watershed data helps in relating the hydrological runoff to the water chemistry. This statistical approach works better in the case of the Turkey Lakes data because the sampling frequencies are relatively higher and the hydrological influences are more significant. In the case of the Risdalsheia data, the sampling is more irregular and the missing gaps are larger. Improvement on the time series simulation of the chemistry data by incorporating the hydrological runoff data is therefore more significant for Turkey Lakes than for Risdalsheia. Thus, for the Turkey Lakes, this statistical technique can be used for estimating the chemical fluxes from the runoff and concentration data. For Risdalsheia, the technique can also be applicable with more frequently sampled data.

ACKNOWLEDGEMENT

This work has been undertaken as a collaborate study between the National Water Research Institute and the Norwegian Computing Center with support through an Environment Canada research contract. The RAIN project is an international collaborative research effort funded by agencies in Norway, Canada, Sweden and U.K. The authors thank Dr. D.S. Jeffries for providing the Turkey Lakes data and Mr. F.C. Elder for his advice and support during all stages of this project.

REFERENCES

Akaike, H. (1973). Maximum likelihood identification of Gaussian auto-regressive moving average models. Biometrika, 60, 255-265.

Akaike, H. (1974). Markovian representation of stochastic processes and its application to the analysis of autoregressive moving average processes. Annals of the Institute of Statistical Mathematics, 26, 363-387.

Akaike, H. (1975). Markovian representation of stochastic processes by canonical variables. SIAM J. Control, 13, 162-173.

Barham, S.Y. and Dunstan, F.O.J. (1982). Missing values in time series. In Time Series Analysis, Theory and Practice. 2. Anderson, O.D. ed. North-Holland, Amsterdam.

Box, G.E.P. and Jenkins, G.M. (1974). Time Series Analysis, Forecasting and Control. Holden Day, San Francisco.

Chen, G.W., Gherini, S.A., Deon, J.O., Hudson, R.J.M. and Goldstein, R.A. (1982). In Ann Arbor Science Book, Schnoor, J.L. ed.

Christophersen, N., Seip, H.M. and Wright, R.F. (1982). A model for streamwater chemistry at Birkenos, Norway. Water Resources Research, 18, 977-996.

Cosby, B.J., Wright, R.F., Hornberger, G.M. and Galloway, J.N. (1984). An equilibrium model for soil and stream water chemistry of White Oak Run, Virginia. Water Resources Research, 20.

Damsleth, E. (1986). Modeling river acidity - A transfer function approach. In Statistical Aspects of Water Quality Monitoring, El-Shaarawi, A.H. and Kwiatkowski, R.E., eds., Elsevier, Amsterdam.

Harvey, A.C. and Pierse, R.G. (1984). Estimating missing observations in economic time series. J. Am. Statist. As., 79, 125-131.

Hipel, K.W. and McLeod, A.I. (1987). Time series modelling for water resources and environmental engineers. Elsevier, Amsterdam.

Jeffries, D.S., Semkin, R.G., Neureuther, R., Seymour, M. and Nicolson, J.A. (1986). Influence of atmospheric deposition on lake mass balance in the Turkey Lakes Watershed, Central Ontario. Water, Air and Soil Pollut., 30, 1033-1044.

Jones, R.H. (1980). Maximum likelihod fitting of ARMA models to time series with missing observations. Technometrics, 22, 389-396.

Kohn, R. and Ansley, C.F. (1986). Estimation, prediction and interpolation for ARIMA models with missing data. J. Am. Statist. As., 81, 751-761.

Lam, D.C.L. and Bobba, A.G. (1984). Modelling watershed runoffs and basin acidification. In Hydrological and Hydrogeochemical Mechanisms and Model Approaches to the Acidification of Ecological Systems. Johansson, I, ed. Nordic Hydrological Programme - report no. 10, Stockholm.

Lam, D.C.L., Boregowda, S., Bobba, A.G., Jeffries, D.S. and Patry, G.G. (1986). Interfacing hydrological and hydrogeochemical models for simulating streamwater chemistry in the Turkey Lakes Watershed, Canada. Water, Air and Soil Pollut., 30, 149-154.

Shibata, R. (1985). Various model selection techniques in time series analysis. In Handbook of Statistics 5. Hannan, E.J. ed., North-Holland, Amsterdam.

Whitehead, P.G., Ned, C., Seden-Perriton, S., Christophersen, N. and Langan, S. (1984). A time series approach to modelling stream activity. In Hydrological and Hydrogeochemical Mechanisms and Model Approaches to the Acidification of Ecological Systems. Johansson, I, ed. Nordic Hydrological Programme - report no. 10, Stockholm.

Wright, R.F., Gjessing, E., Christophersen, N., Lotse, E., Seip, H.M., Semb, A., Sletaune, B., Storhaug, R. and Wedum, K. (1986). Project RAIN: changing acid deposition to whole catchments. The first year of treatment. Water, Air and Soil Pollut., 30, 47-63.

APPENDIX A. Kalman filter representation of non-stationary ARIMA-models for time series with missing observations.

A.1 INTRODUCTION

It has long been recognized that the best (only?) way to solve the problem of estimating the parameters in an ARMA-model for a time series with missing observations is to formulate the model as a Dynamic Linear Model (DLM) and apply the Kalman filter, which handles the missing observations automatically, and in an optimal way. Jones (1980) gives an excellent presentation. Jones approach, however, handles only the stationary situation. Harvey & Pierse (1984) give one generalization to non-stationary models, but their approach requires the first, or last, d+1 observations in the series to be observed, where d is the number of differences required to obtain stationarity. Kohn & Ansley (1986) give a general solution, by introducing an extended version of the Kalman filter, so that standard algorithms and programs for Kalman filtering are not applicable. In this appendix we generalize the approach of Jones (1980) to the non-stationary case, without any requirements on the number of observations in the beginning or end of the series.

A.2 MODEL FORMULATION AND THE STATIONARY ALGORITHM

We assume that the time series $\{x_t\}$ follows an ARIMA (p, d, q) model given by

$$\phi(B)(1 - B)^d (x_t - \mu) = \theta(B) a_t \qquad (A.1)$$

where B is the backwards shift operator so that $B_k x_t = x_{t-k}$,

$$\phi(B) = 1 - \phi_1 B - \ldots - \phi_p B^p$$

$$\theta(B) = 1 - \theta_1 B - \ldots - \theta_q B^q$$

are polynomials in B with all their roots outside the unit circle, d is the number of differences necessary to make $(1-B)^d x_t$ stationary, and μ is the level of the series if d=0. If d>0, μ=0. Except for a few changes in notation, the following is based on Jones (1980). As shown by Akake (1973, 1974, 1975), the ARIMA model (A-1) above can be given a Markovian representation as follows:

The state of the process is a vector of dimension m = max (p+d, q+1) given by

$$z(t) = (x(t|t), x(t+1|t), \ldots , x(t+m-1|t))' \qquad (A.2)$$

where $x(t+j|t)$ denotes the projection of x_{t+j} on the values of the time series up to and including t. $x(t+j|t)$ is thus the j-step prediction from time t, and $x(t|t) = x_t$; the value of the process at time t.

Let the matrix F and vector G be given by

$$F = \begin{bmatrix} 0 & 1 & 0 & \dots & & 0 \\ 0 & 0 & 1 & 0 & \dots & 0 \\ & & \cdot & & & \\ & & \cdot & & & \\ & & \cdot & & & \\ 0 & 0 & \dots & 0 & 0 & 1 \\ \alpha_m & \alpha_{m-1} & \dots & & \alpha_2 & \alpha_1 \end{bmatrix} \qquad G = \begin{bmatrix} 1 \\ \psi_1 \\ \cdot \\ \cdot \\ \cdot \\ \psi_{m-1} \end{bmatrix} \qquad (A.3)$$

where the α_i are given by

$$(1 - \alpha_1 B - \dots - \alpha_{p+d} B^{p+d}) = (1 - B)^d \phi(B) \qquad (A.4)$$

$$\alpha_i = 0, \qquad p+d < i \leq m$$

and the ψ_i are the coefficients in the polynomial $\psi(B)$ given by

$$\psi(B) = (1 - B)^{-d} \phi^{-1}(B)\theta(B) \qquad (A.5)$$

Then we have

$$z(t + 1) = Fz(t) + Ga_{t+1} \qquad (A.6)$$

In the framework of dynamic linear models (DLM) and Kalman-filtering, (A.6) is known as the equation of state. The associated observational equation is given by

$$x(t) = Hz(t) + v(t) \qquad (A.7)$$

with $H = (1, 0, \ldots, 0)$ and where $v(t)$ denotes (possible) observation errors which are uncorrelated at different times and which are uncorrelated with the noise series. We assume $Ev(t)=0$, $\mathrm{Var}\ v(t)=R$, where $R=0$ if there are no observation errors. When the model is formulated in this way, it is straightforward to calculate the innovations and their variance recursively, using the Kalman filter. The procedure is outlined in detail in Jones (1980), and works just as well when there are missing observations in the series. From the innovations and their variances, the exact likelihood can be calculated for any stationary ARMA-model with any pattern of missing observations. The algorithm can briefly be summarized as follows:

Define $z(t+j|t)$ as the projection of $z(t+j)$ on the observations up to and including time t, so that $z(t|t) = z(t)$. Further define $P(t+j|t)$ as the covariance matrix of $z(t+j|t)$. The algorithm consists of four steps:

1. $$z_i(0|0) = \mu,\quad P_{ij}(0|0) = \gamma_{|i-j|} - \sum_{k=0}^{\min(i,j)-1} \psi_k\, \psi_{k+|i-j|}$$

where the γ_i denotes the covariances of the process, $\gamma_i = \mathrm{cov}(x_t, x_{t-i})$

Repeat steps 2 and 3 for $t = 1, \ldots, n$.

2. $z(t+1|t) = F\, z(t|t)$

   $P(t+1|t) = F\, P(t|t)\, F' + GG'$

3. $\Delta(t+1) = P(t+1|t)\, H'[H\, P(t+1|t)\, H' + R]^{-1}$

   $\tilde{x}(t+1) = x(t+1) - x(t+1|t)$

   $z(t+1|t+1) = z(t+1|t) + \Delta(t+1)\, \tilde{x}(t+t)$

   $P(t+1|t+1) = P(t+1|t) - \Delta(t+1)\, H\, P(t+1|t)$

   $v_{t+1} = P_{11}(t+1|t) + R$

Calculate the $-2\ell n$ likelihood ℓ by

4. $\ell = \sum_{t=1}^{n} [\ell n\, \sigma^2\, v_t + \tilde{x}_t^2/\sigma^2\, v_t]$

In the case of missing observations step 3 in the algorithm is replaced by

3b. $z(t+1|t+1) = z(t+1|t)$

   $P(t+1|t+1) = P(t+1|t)$

and the corresponding terms are skipped in step 4.

A.3 NON-STATIONARY CASE

The formulation of the model and the algorithm from the previous section can be applied in the non-stationary case as well, except that Step 1 in the algorithm breaks down, since the convariances y_i of the process do not exist when the process is non-stationary. When there are no missing observations the straightforward solution is to difference the series d times, and apply the scheme above to the differenced, stationary series.

However, when there are missing observations in the series, this method does not work, since taking differences over varying time spans will introduce a covariance structure which will depend on time. Alternatively, we could define $w_t=(1-B)^d x_t$ as missing if at least one of x_t, x_{t-1}, x_{t-d} were missing. This will work satisfactorily if there are only a few missing values, but when there is a substantial number of observations missing this will result in a serious loss of information. The beauty of the Kalman filter approach is, among others, its ability to handle missing observations effectively, and we would like to utilize this ability in the non-stationary case as well.

In the non-stationary case, the state vector $z(t|t)$ will be non-stationary until at least d observations are actually observed. Let k_1, k_2, ..., k_d be the times of the first d actually observed values of the series, $k_i \geq i$. The first stationary state vector is then

$$z(k_d|k_d) = (x(k_d|k_d),\ x(k_d+1|k_d),\ \ldots\ ,\ x(k_d+m-1|k_d))'$$

If we can find an expression for $z(k_d|k_d)$, and the corresponding $P(k_d|k_d)$, we can use these as starting values, and run the Kalman algorithm from $t=k_d+1$ on.

A3.1 The State Vector

Let $w_t = (1-B)^d x_t$, $t = \ldots, -1,0,1,\ldots$ be the stationary series which generates x_t by summation d times. Consider the jth element in $z(k_d|k_d)$, $x(k_d+j-1|k_d)$, which is the projection of x_{k_d+j-1} on the observations up to and including k_d. We then have, for j>1,

$$w_{k_d+j-1} = (1-B)^d \, x_{k_d+j-1} =$$

$$\frac{(1-\beta_1^{(j)} B^{j-1} - \beta_2^{(j)} B^{k_d-k_{d-1}+j-1} - \ldots - \beta_d^{(j)} B^{k_d-k_1+j-1})}{(1-\delta_1^{(j)} B - \ldots - \delta_{k_d-k_1+j-1-d}^{(j)} B^{k_d-k_1+j-1-d})} \, x_{k_d+j-1} \qquad (A.8)$$

The β's and δ's can be found by solving the linear equation system obtained by equating coefficients in

$$(1-B)^d \, \delta^{(j)}(B) = \beta^{(j)}(B) \qquad (A.9)$$

where $\delta^{(j)}(B)$ and $\beta^{(j)}(B)$ are respectively the denominator and numerator in the right hand side of (A.8). Multiplying both sides of (A.8) and rearranging, we obtain

$$x_{k_d+j-1} = \sum_{i=1}^{d} \beta_i^{(j)} x_{k_d-i+1} + w_{k_d+j-1} - \sum_{i=1}^{j-1} \delta_i^{(j)} w_{k_d+j-1-i}$$

$$- \sum_{i=j}^{k_d-k_1+j-d} \delta_i^{(j)} \, w_{k_d+j-1-i} \qquad (A.10)$$

Projecting both sides of (A.10) on all available information up to and including time k_d, we obtain

$$\begin{aligned} x(k_d|k_d) &= x_{k_d} \\ x(k_d+j-1|k_d) &= \sum_{i=1}^{d} \beta_i^{(j)} x_{k_d+1-i} \end{aligned} \qquad (A.11)$$

where we have set the unobserved values of w_t to their mean, 0, for $t < k_d$, and where we have utilized $w(k_d+j-1|k_d)=0$.

From (A.11) we obtain

$$z(k_d|k_d) = B\ (x_{k_d}, x_{k_{d-1}}, \ldots, x_{k_1})' \qquad (A.12)$$

where

$$B = \begin{bmatrix} 1 & 0 & 0 & \ldots & 0 \\ \beta_1^{(2)} & \beta_2^{(2)} & & \ldots & \beta_d^{(2)} \\ & & \cdot & & \\ & & \cdot & & \\ & & \cdot & & \\ \beta_1^{(m)} & \beta_2^{(m)} & & \ldots & \beta_d^{(m)} \end{bmatrix} \qquad (A.13)$$

From (A.10) we can write

$$P(k_d|k_d) = \mathrm{cov}(z(k_d|k_d)) = \mathrm{cov}\ (D_1 w(k_d|k_d) + D_2\ \bar{w}_{k_d}) \qquad (A.14)$$

where D_1 is the mxm matrix given by

$$D_1 = \begin{bmatrix} 0 & 0 & 0 & \cdots & & 0 \\ -\delta_1^{(2)} & 1 & 0 & \cdots & & 0 \\ -\delta_2^{(3)} & -\delta_1^{(3)} & 1 & & & 0 \\ & & \vdots & & & \\ -\delta_{m-1}^{(m)} & -\delta_{m-2}^{(m)} & & \cdots & -\delta_1^{(m)} & 1 \end{bmatrix} \qquad (A.15)$$

D_2 is the $m x(k_d - k_1 - d)$ matrix given by

$$D_2 = \begin{bmatrix} 0 & \cdots & 0 \\ -\delta_2^{(2)} & \cdots & -\delta_{k_d-k_1-d+1}^{(2)} \\ -\delta_3^{(3)} & \cdots & -\delta_{k_d-k_1-d+2}^{(3)} \\ & \vdots & \\ -\delta_m^{(m)} & \cdots & -\delta_{k_d-k_1-d+m-1}^{(m)} \end{bmatrix} \qquad (A.16)$$

and $\bar{w}_{k_d} = (w_{k_d-1}, w_{k_d-2}, \ldots, w_{k_1+d})'$. Then from (A.14),

$$P(k_d | k_d) = D_1 P_w(0|0) D_1' + D_1 \Pi D_2' + D_2 \Pi D_1' + D_2 \Gamma D_2' \qquad (A.17)$$

where $P_w(0|0)$ is the mxm initial covariance matrix of the state vector for the stationary part of the model, Π is the $m\, x(k_d - k_1 - d)$ matrix $E(w(k_d | k_d) \cdot \bar{w}_{k_d}')$ with elements given by $\Pi_{ij} = \psi_{i+j-1}$ and Γ is the $(k_d - k_1 - d) x(k_d - k_1 - d)$ covariance matrix of the stationary part of the process, with elements $\Gamma_{ij} = \gamma_{|i-j|}$.

Example :

Suppose $d=2$, and assume $k_1=1$ without loss of generality. Let $k_2=k$ to simplify notation. Then, from (A.9)

$$(1-2B+B^2)(1-\delta_1^{(j)} B - \dots - \delta_{k-4+j}^{(j)} B^{k-4+j}) = 1-\beta_1^{(j)} B^{j-1} - \beta_2^{(j)} B^{k-2+j}, \quad j=2, \dots, m$$

which, after some calculations, gives

$$\beta_1^{(j)} = 1 + (j-1)/(k-1), \quad \beta_2^{(j)} = -(j-1)(k-1)$$

$$\delta_i^{(j)} = -(i+1), \; i \leq j-2, \; = -(j-1)\left(1 - \frac{i+2-j}{k-1}\right), \; i=j-1, \dots, k-4+j$$

Thus $x(k+j|k) = x_k + (x_k - x_1)(j-1)/(k-1), \quad j=1, \dots, m$

Let us further assume that $m=3$. Then

$$D_1 = \begin{bmatrix} 0 & 0 & 0 \\ \frac{k-2}{k-1} & 1 & 0 \\ \frac{2(k-2)}{k-1} & 2 & 1 \end{bmatrix}, \quad D_2 = \begin{bmatrix} 0 & 0 & \dots & 0 \\ \frac{k-3}{k-1} & \frac{k-4}{k-1} & \dots & \frac{1}{k-1} \\ \frac{2(k-3)}{k-1} & \frac{2(k-4)}{k-1} & \dots & \frac{2}{k-1} \end{bmatrix}$$

If the model, for example, is an ARIMA (0,2,1) then

$$P_w(0|0) = \begin{bmatrix} 1+\theta^2 & -\theta & 0 \\ -\theta & \theta^2 & 0 \\ 0 & 0 & 0 \end{bmatrix} \qquad \Pi = \begin{bmatrix} -\theta & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \end{bmatrix}$$

$$\Gamma = \begin{bmatrix} 1+\theta^2 & -\theta & 0 & \dots & 0 \\ -\theta & 1+\theta^2 & -\theta & \dots & 0 \\ & & \vdots & & \\ 0 & \dots & 0 & -\theta & 1+\theta^2 \end{bmatrix}$$

which, using (A.17) and simplifying, gives

$$P(k|k) = [\theta^2 + \frac{k-2}{6(k-1)} (2k-3 - 4k\theta + (2k-3)\ \theta^2)] \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 2 \\ 0 & 2 & 4 \end{bmatrix}$$

Thus, in this case, the initial convariance matrix is singular, and the variances increase approximately linearly in k.

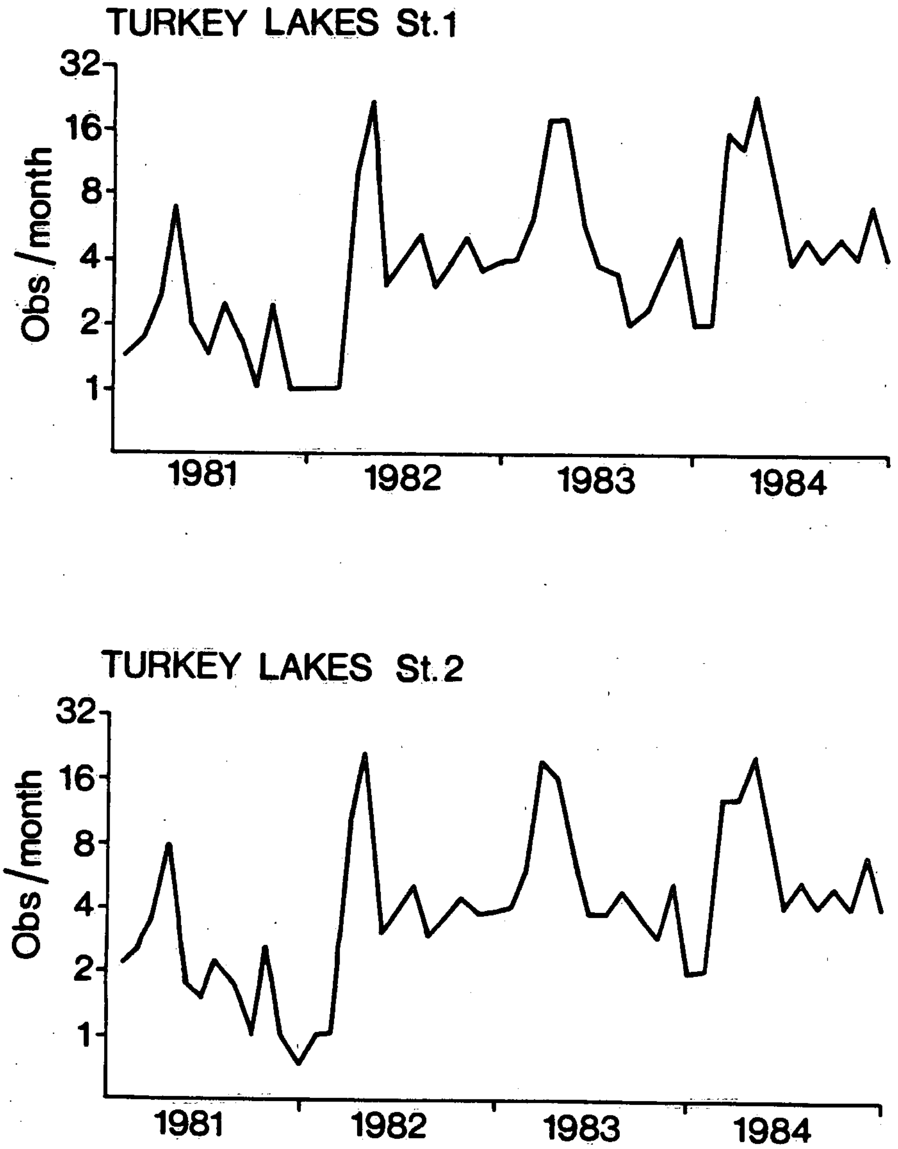
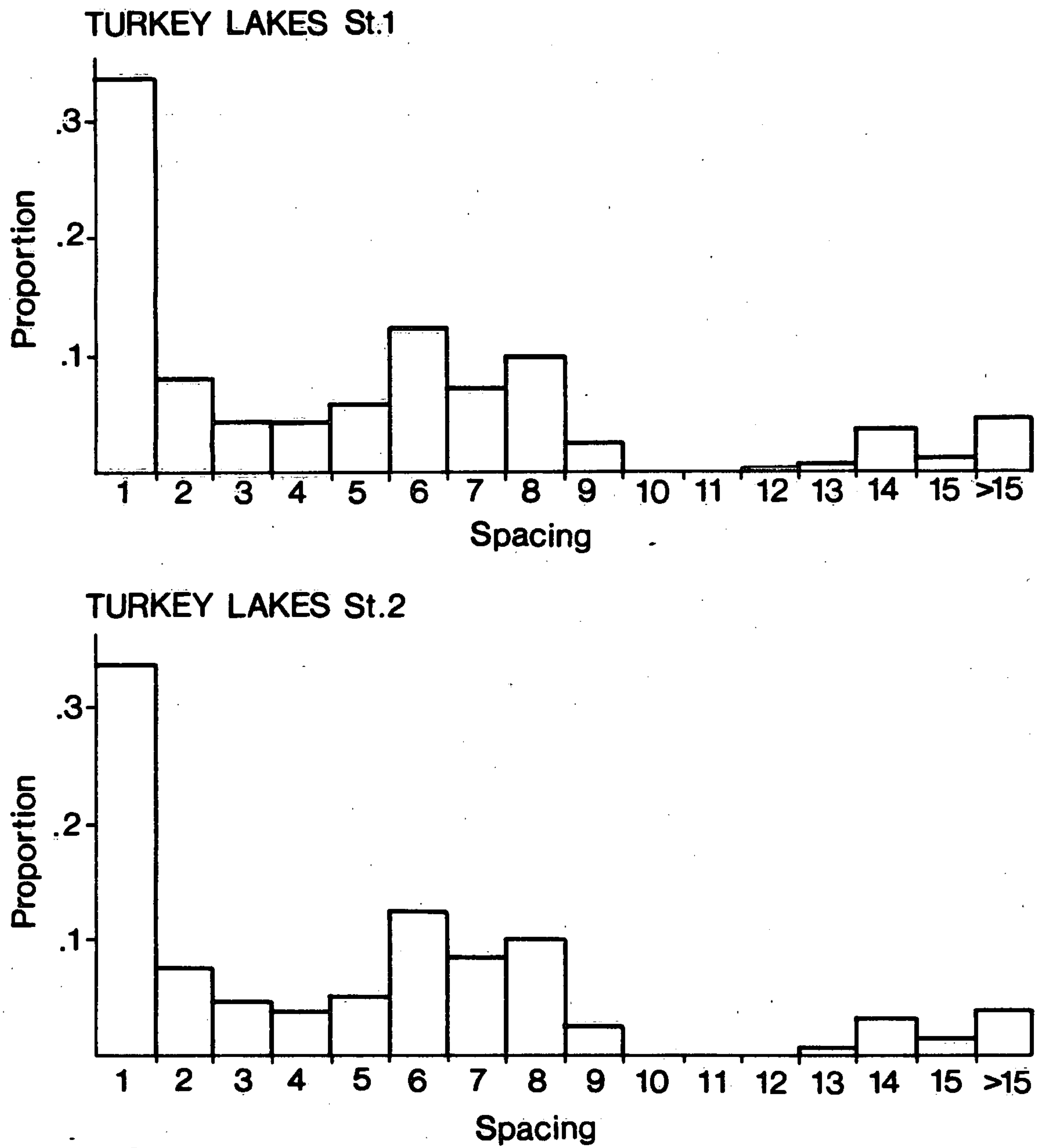

TURKEY LAKES St.1
Obs /month
32
16
8
4
2
1
1981
1982
1983
1984
TURKEY LAKES St.2
Obs /month
32
16
8
4
2
1
1981
1982
1983
1984

FIGURE 4.1



TURKEY LAKES St.1
Proportion
.3
.2
.1
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
>15
Spacing
TURKEY LAKES St.2
Proportion
.3
.2
.1
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
>15
Spacing

FIGURE 4.2

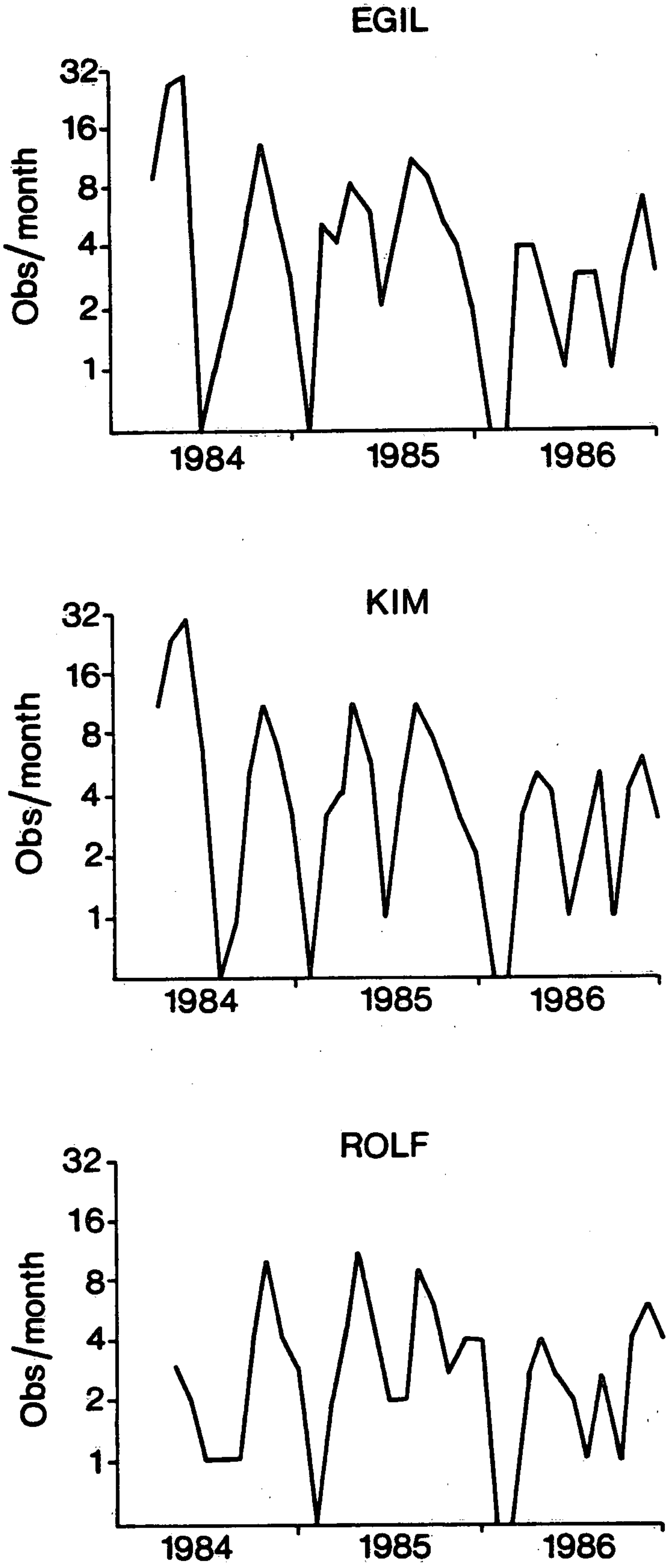

EGIL
32
16
8
4
2
1
Obs/month
1984
1985
1986
KIM
32
16
8
4
2
1
Obs/month
1984
1985
1986
ROLF
32
16
8
4
2
1
Obs/month
1984
1985
1986

FIGURE 4.3

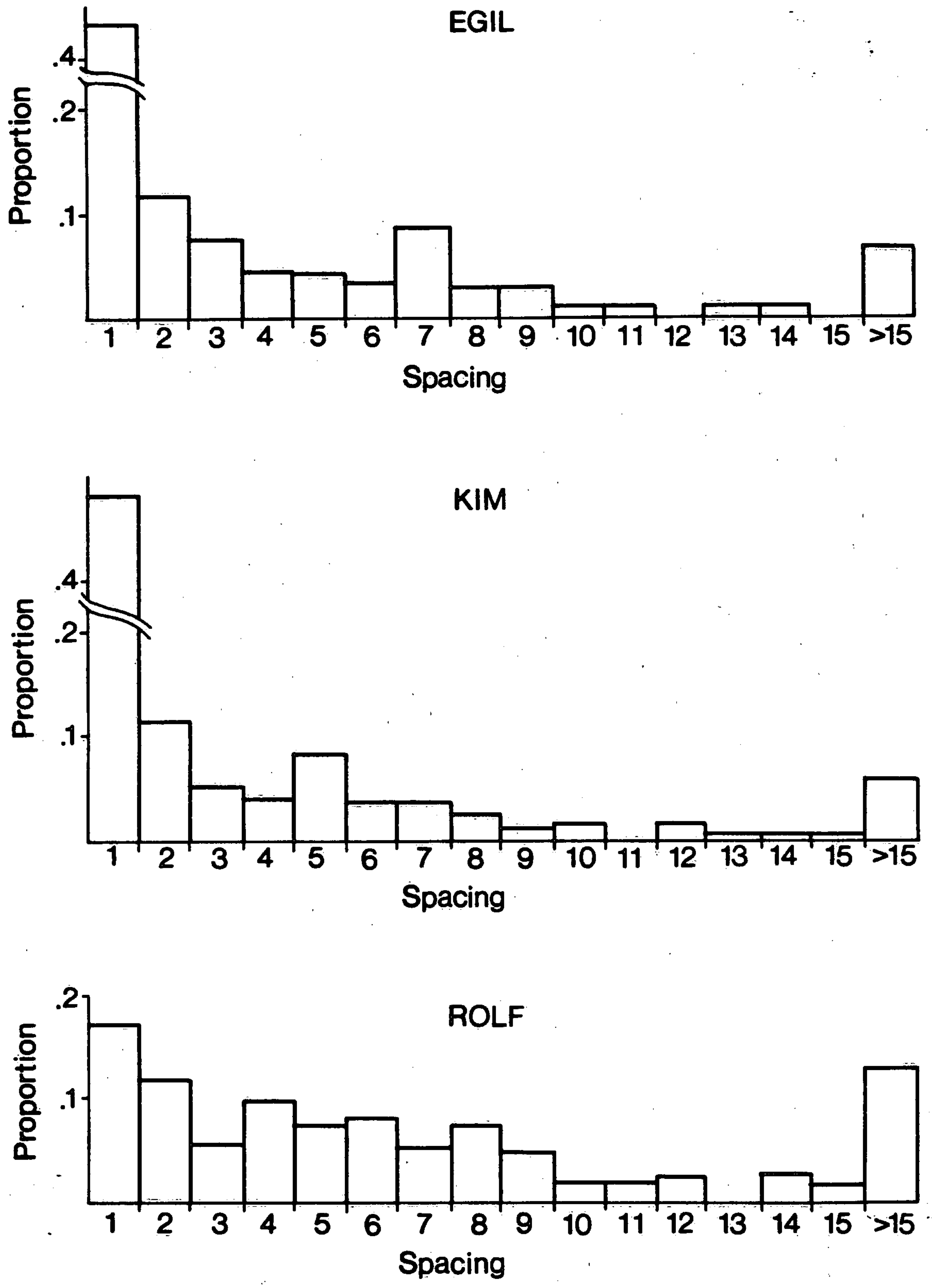

EGIL
Proportion
.4
.2
.1
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
>15
Spacing
KIM
Proportion
.4
.2
.1
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
>15
Spacing
ROLF
Proportion
.2
.1
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
>15
Spacing

FIGURE 4.4