A dynamic factor model for forecasting surface water quality

Yun-Sheng Yu, Shimin Zou

Department of Civil Engineering, The University of Kansas, Lawrence, KS 66045, USA

ABSTRACT

A dynamic factor model and an autoregressive model for factor scores are developed for surface water-quality data. The combined model permits the state-space formulation of the system and the use of the Kalman filter for forecasting the time series of water quality constituents. The calibrated model using the monthly water quality data for ten constituents at Station 214 of the Neosho River is used for forecasting the time series of water quality constituents.

Introduction

Measurements of water quality constituents at a monitoring station in a watershed represent the cumulative effects of all sources of pollutants upstream from the station. These constituents do not respond instantaneously to the changes in land uses, climate, and sources of pollutants, etc. in the watershed. Their time-lagged effects play an important role in modeling water quality time series. Although the traditional factor analysis has been widely accepted in diverse fields (Anderson [3], Akaike [1]). its lack of a proper way to handle a lagged covariance structure makes itself undesirable for use in water quality data analysis (Anderson [2]). In addition, the traditional factor model is time-invariant and can not be used for forecasting. The purpose of this paper is to extend the dynamic factor model developed by Zou and Yu [10] for forecasting one-month ahead concentrations of surface water quality constituents.

A General Dynamic Factor Model

Let $X(t)$ be a mean-centered p vector-valued second-order stationary time series of water quality constituents with
covariance function $C(u)$ where $u=0,1,2,...$ is the time-lag. The general dynamic factor model (Molenaar [7]) has the following form

$$X(t) = \sum_{u=0}^{s} \Lambda(u) \eta(t-u) + \epsilon(t), t=0, \pm 1, \ldots \quad (1)$$

where the lag $s$ is an unknown positive integer; $\Lambda(u)$, for $u=0, 1, \ldots, s$, are a sequence of $(p \times q)$ dimensional lagged loading matrices, $\eta(t)$ are a $q$ vector-valued latent factor series; $\epsilon(t)$ are $p$ vector-valued random errors; $q$ is always less than $p$. The assumed stationarity of $X(t)$ generally will give rise to a decaying filter as $u$ increases (Hannan [4]) and $\Lambda(u)$ can be truncated at a suitable finite lag, such as $u = s$. The dynamic factor model given by Eq.1 covers the entire covariance function of $X(t)$:

$$C(u) = \sum_{v=0}^{s} \sum_{w=0}^{s} \Lambda(v) \Phi(u+w-v) \Lambda'(w) + \Theta(u), \quad (2)$$

If the number of factors $q$ and the number of lags are chosen, then the dynamic factor model can be rewritten as a system of simultaneous structural equations as follows. For an arbitrary time $t$, covering a subset of $a+1$ consecutive time-points $t-a,t-a+1, \ldots, t$, where $a \geq s$, from Eq.2, one can obtain

$$C(u) = \Lambda \Phi \Lambda' + \Theta \quad (3)$$

where

$$\Phi = \{ \Phi(i-j); i,j=1, \ldots, a+s+1 \}, \quad (4)$$

and

$$\Theta = \{ \Theta(i-j); i,j=1, \ldots, a+1 \} \quad (5)$$

are covariance matrices.

In general, $\Phi(u)$ represents a sequence of $(q \times q)$ dimensional matrices of lagged covariances and $\Theta(u)$ is a sequence of $(p \times p)$ dimensional matrices of lagged covariances. For the system to be identifiable, constraints must be imposed upon $\Phi(u)$. The following less restrictive constraints are proposed for parameter estimation:

$$\Phi(u) = \begin{cases} I_q & u = 0 \\ \text{diag} \left( \phi_{u1}, \ldots, \phi_{uq} \right) & 1 \leq u \leq s \\ 0 & u > s \end{cases} \quad (6)$$

In Eq.3, $C$ and $\Theta$ are $[p(a+1) \times p(a+1)]$ dimensional covariance matrices; $\Lambda$ is a $[p(a+1) \times q(a+s+1)]$ matrix of factor loadings; $\Phi$ is a $[q(a+s+1) \times q(a+s+1)]$ matrix of factors and $\Theta$ represents the vector of unknown parameters. In application, the matrix $C$ in Eq.3 can be replaced by the sample covariance matrix estimated from the original data set. The right side of Eq.3 must be
estimated by using either the method of maximum likelihood or the least-squares method (Yu [8]) depending on prior information available. The total number of known elements in Eq.3 is \( N_\text{c} = \frac{1}{2} p(p+1) + ap^2 \), the total number of unknown elements to be estimated is \( N_\theta = pq(s+1) + qs + p \), and the degree of freedom of the system is \( df = N_\text{c} - N_\theta \). The dimensions of the matrices involved in Eq.3 depends on the values of \( p, q, s, \) and \( a \).

**Surface Water Quality Data and Model Identification**

Monthly data of ten water quality constituents at Station 214 on the Neosho River in Kansas from December 1975 through November 1989 were selected for model identification. These constituents are: alkalinity (ALKAL), ammonia plus organic nitrogen (AMORG), chloride (CHLOR), dissolved solids (DISLD), dissolved oxygen (DISOX), potassium (PTSUM), specific conductance (SPCND), sulfate (SULFT), total hardness (TOHRD) and total phosphorus (TOPHS). These variables were standardized with zero mean and unity variance before modeling. The sample correlation matrices \( C(0), C(1), \) and \( C(2) \) for lag zero, one, and two, presented elsewhere (Zou [9]), show that the absolute values of matrix elements generally decrease as the time lag increases as expected. It is thus reasonable to neglect correlation matrices beyond lag two in modeling monthly water quality time series. Thus four different models are identified. They are \( M_0 \) with \( a=s=0 \), \( M_1 \) (\( a=s=1 \)), \( M_2 \) (\( a=2 \) and \( s=1 \)), and \( M_3 \) (\( a=s=2 \)). For the \( M_0 \) model, the computed factor loading matrix for

| Table 1 | Factor Loading matrix \( \Lambda(0) \) for the first three factors and specific variance \( \Theta(0) \) |
|---------|---------------------------------|-----------------|-----------------|-----------------|
|         | Factor 1 | Factor 2 | Factor 3 | Specific Variance |
| ALKAL   | 0.736     | -0.478   | -0.191   | 0.194            |
| AMORG   | -0.331    | 0.119    | 0.280    | 0.798            |
| CHLOR   | 0.718     | 0.175    | -0.318   | 0.352            |
| DISLD   | 0.983     | 0.117    | 0.055    | 0.016            |
| DISOX   | 0.097     | -0.458   | 0.075    | 0.775            |
| PTSUM   | -0.006    | 0.655    | -0.333   | 0.460            |
| SPCND   | 0.974     | 0.038    | 0.015    | 0.050            |
| SULFT   | 0.791     | 0.462    | 0.403    | 0.001            |
| TOHRD   | 0.962     | -0.141   | 0.068    | 0.050            |
| TOPHS   | -0.520    | 0.133    | -0.000   | 0.712            |

the first three factors and the specific variances are shown in Table 1. The first three factors are selected because the eigenvalue of the fourth factor is 0.390 much
less than one and its incremental contribution to the cumulative percentage of explainable variance is 3.9%, a relatively small value.

**Calibration of the M\(_1\) Model**

Based on the criterion of overall goodness-of-fit of the model and the principle of parsimony (Zou [9]), the M\(_1\) model, which includes correlation matrices C(0) and C(1), was chosen among the four different models identified. The M\(_1\) model with p=10 variables, q=3 factors, and a=s=1 includes the following unknown coefficient matrices in Eq. 3: θ(0), Λ(0), Λ(1), and $Φ(1)$, where Λ(0) and Λ(1) are (10×3) factor loading matrices; θ(0) is a (10×10) diagonal matrix; Φ(1) is a (3x3) diagonal matrix. The degree of freedom of this system is 82 and the total unknown elements are 73.

The first 13 years of monthly data of the ten constituents were used for estimating the optimal parameters in Eq.3, namely, Λ(0), Λ(1), $Φ(1)$, and θ(0), to minimize the sum of the squared residual errors, i.e. the differences between corresponding elements in the computed and the sample covariance matrices C(0) and C(1). For the M\(_1\) model, this value is equal to 0.805 and the model is considered acceptable.

**A Multivariate Autoregressive Model for Dynamic Factor Vector**

The following multivariate autoregressive model for the dynamic factor vector $η(t)$ is proposed

$$η(t) = G_1 η(t-1) + G_2 η(t-2) + α(t-1) \quad (7)$$

where $G_1$ and $G_2$ are coefficient matrices and $α(t-1)$ is a vector-valued white noise series with zero mean and covariance matrix $Q = \text{diag}(q_1, ..., q_q)$. Matrices $G_1$, $G_2$, and $Q$ in Eq.7 are estimated by using the method of moments to preserve explicitly a limited number of moments of the original time series of factor scores. The detailed derivation is presented elsewhere (Zou [9]).

**Kalman Filtering**

Equations 1 and 7 can be put into the following state-space form in a straightforward manner:

$$Y_{t+1} = H_t Y_t + G_t η_t \quad (8)$$

and

$$X_{t+1} = S_{t+1} Y_{t+1} + ε_{t+1} \quad (9)$$

Equations 8 and 9 are, respectively, the state transition
equation and the measurement equation in the state-space form. \( H_t, S_t, \) and \( G_t \) are, respectively, the transition matrix, the measurement matrix, and the noise transition matrix.

With Eqs. 8 and 9 and the Kalman [5] filter algorithm, optimal estimates of state variables at one-month-ahead can be obtained recursively when the system matrices \( (S, G, H) \), system noise statistics \( (Q) \), measurement noise statistics \( (R=\Theta(0)) \) and initial states are known. The Kalman filter equations (Lewis [6]) may be summarized without derivation as follows:

1). The state prediction:
\[
\hat{Y}_{t+1} = H\hat{Y}_t \quad (10)
\]

2). Error covariance of predicted state:
\[
\hat{P}_{t+1} = HP_tH' + GQ_tG' \quad (11)
\]

3). Kalman gain factor:
\[
K_{t+1} = \hat{P}_{t+1}S'\left(S\hat{P}_{t+1}S' + R_{t+1}\right)^{-1} \quad (12)
\]

4). Updated state:
\[
\tilde{Y}_{t+1} = \hat{Y}_{t+1} + K_{t+1}(X_{t+1} - SY_{t+1}) \quad (13)
\]

5). Error covariance matrix of updated state:
\[
\hat{P}_{t+1} = \hat{P}_{t+1} - K_{t+1}SP_{t+1} \quad (14)
\]

where the symbol "\(^T\)" indicates matrix transpose, the symbol "\(^\ast\)" designates the estimated value after an observation has been made, and the symbol "\(~\)" denotes the estimated value at time \( t+1 \) conditioned on observations up to time \( t \).

It is assumed that \( H, G, X, \) and \( S \) are known for all time \( t \). Also, the variance matrix of the estimated state vector \( \hat{Y}_t \) is denoted by \( \hat{P}_t \). Thus the algorithm estimates recursively the state vector \( Y_{t+1} \) and its covariance matrix \( \hat{P}_{t+1} = \text{var}(\tilde{Y}_{t+1}) \).

**Forecasting Time Series of Water Quality Constituents**

The 14-year records for the ten water quality constituents at Station 214 on the Neosho River were divided into two parts. The first 13-year data were used for model calibration and and the last year data for comparing with the forecast values. The measured variables were standardized with zero mean and unity variance. The calibrated model together with the Kalman filter was used to compute the monthly values of the 14th year. These values are compared with the actual measurements. The model errors as measured by the residual error variances between the computed values and the measurements for the ten water quality constituents
are, respectively: 0.271, 0.231, 0.330, 0.002, 0.598, 0.467, 0.070, 0.168, 0.016 and 0.488, in the order of the constituents listed in Table 1. The computed and measured 14-year monthly time series for three standardized variables, namely, dissolved solids, sulfate, and potassium are shown in Figs.1,2, and 3, respectively. The monthly values of the 14th year are the forecast values. These three constituents were chosen to illustrate the effects of residual error variances of the model on different constituents. As expected, the computed time series for dissolved solids with residual error variance of 0.002 agrees extremely well with the observed time series as shown in Fig.1. In other words, the first three factors are sufficient for modeling this constituent. The residual error variance for sulfate is 0.168 and the agreement between the computed and measured values is still reasonably good as shown in Fig.2. However, for potassium, the computed time series generally underestimates the standardized concentrations as shown in Fig.3 because of the relatively large residual error variance of 0.467. The results can be improved by either including additional factors or by imposing an upper bound on the specific variance for the selected constituent in parameter estimation.

Conclusions

Measurements of water quality constituents represent the cumulative effects of all sources of pollutants upstream from the monitoring station in the watershed. The time delayed effects due to changes in land uses, sources of pollution, and climate, etc. must be included in modeling the water quality time series. This paper developed a dynamic factor model and an multivariate autoregressive model for the factor scores to account for the time-lagged covariance matrices. The combined model and the Kalman filter are then used for modeling and forecasting the time series of surface water quality constituents. As an example, the 14-year records for ten constituents at Station 214 of the Neosho River in Kansas were used for model calibration and forecasting. For constituents with small residual error variances, the computed time series generally agree well with the measured values. On the other hand, for constituents with relatively large residual error variances, the model generally underestimates the measured values of the standardized variables.

References

9. Zou, Shimin. 'Trend Analysis and Modeling of Water Quality of the Kansas River and the Arkansas River' Ph. D. dissertation, Department of Civil Engineering, the University of Kansas, Lawrence, Kansas, April 1992.
Fig. 2 Time Series of Sulfate

Fig. 3 Time Series of Potassium