기술강좌

수문통계학의 기초(Ⅴ)

Introduction to Statistical Hydrology(Ⅴ)

허 준 행*

머리말

지난 강좌까지는 여러 가지 통계값 등 통계학의 기초, 수문자료의 빈도해석 방법, 여러 가지 검정방법, 수문자료의 변동 및 경향 분석방법에 대하여 연자료 및 계절자료인 경우에 대하여 설명하였다. 이번 강좌는 수문통계학의 기초 마지막 편으로 수문자료 해석시 자주 보게 되는 결측치의 보완방법과 자료확충 방법에 대하여 설명하기로 한다.

5. 결측치보완 및 자료확충방법

일반적으로 수문 및 수질자료에서 결측치를 갖거나 자료기간이 비교적 짧은 경우를 많이 볼 수 있다. 이런 자료를 이용하여 해석하는 경우에 편차나 표본오차는 크게 증가된다. 이러한 문제점은 대상지역 주변의 자료를 이용함으로써 경감될 수 있으며, 결측치가 없이 비교적 긴 자료를 갖고 있는 주변 한 지점의 자료를 이용하는 간단한 방법에서부터 주변 여러 지점의 자료를 이용하는 보다 복잡한 방법에 이르기까지 다양한 방법이 있다.

5.1 일반적인 방법

5.1.1 선형회귀방정식

2개의 지점 사이의 정보를 이용하는 간단한 방법으로 선형회귀방정식을 들 수 있다. 2지점의 자료를 각각 Y와 X라 하고, Y의 자료 기간을 N_1, X의 자료 기간을 N_1+N_2라 하면 다음 식과 같이 쓸 수 있다.

$$y_t=\hat{\mu}_{y_1}+\hat{\rho}_{xy}\frac{\hat{\sigma}_{y_1}}{\hat{\sigma}_{x_1}}(x_t-\hat{\mu}_{x_1}) \qquad (5.1)$$

여기서 $\hat{\mu}_{y_1}$, $\hat{\mu}_{x_1}$, $\hat{\sigma}_{y_1}$, $\hat{\sigma}_{x_1}$은 자료수 N_1에 대한 Y와 X의 평균 및 표준편차이고, $\hat{\rho}_{xy}$는 상관수이다. 그러므로, 동시자료(concurrent observation)N_1에 대한 식(5.1)을 이용하여 Y자료를 확충할 수 있다. 확충된 자료를 포함한 자료 Y의 평균과 분산은 다음과 같이 주어진다(Fiering, 1963 ; Matalas와 Jacobs, 1964).

$$\hat{\mu}_y=\frac{N_1\hat{\mu}_{y_1}+N_2\hat{\mu}_{y_2}}{N_1+N_2}$$

$$=\hat{\mu}_{y_1}+\frac{N_2}{(N_1+N_2)}\hat{b}(\hat{\mu}_{x_2}-\hat{\mu}_{x_1}) \qquad (5.2)$$

* 연세대학교 토목공학과 조교수

$$\hat{\sigma}_y = \frac{1}{(N_1+N_2-1)}\left[(N_1-1)\hat{\sigma}_{y_1}+(N_2-1)\hat{b}^2\hat{\sigma}_{x_2}^2 + \frac{N_1N_2}{(N_1+N_2)}\hat{b}^2(\hat{\mu}_{x_2}-\hat{\mu}_{x_1})^2\right] \quad (5.3)$$

여기서 $\hat{\mu}_{x_2}$, $\hat{\sigma}_{x_2}$는 각각 자료수 N_2에 대한 X자료의 평균과 분산값이고, $\hat{b}=\hat{\rho}_{xy}\hat{\sigma}_{y_1}/\hat{\sigma}_{x_1}$ 이다. 식(5.3)의 분산값은 편의된 값으로 이러한 문제점을 해결하기 위하여 식(5.1)을 다음과 같이 수정하여 사용할 수 있다(Matalas와 Jacobs, 1964).

$$y_t = \hat{\mu}_{y_1}+\hat{\rho}_{xy}\frac{\hat{\sigma}_{y_1}}{\hat{\sigma}_{x_1}}(x_t-\hat{\mu}_{x_1}) + \alpha\theta\sqrt{1-\hat{\rho}_{xy}^2}\,\hat{\sigma}_{y_1}\varepsilon_t \quad (5.4)$$

여기서 θ는 지표변수(indicator parameter)로 $\theta=1$이면 오차항을 포함하고, $\theta=0$이면 오차항을 포함하지 않는다. 또한 ε는 오차항(noise term)으로 표준정규분포(평균=0, 분산=1)이며 α는 다음 식과 같이 주어진다.

$$\alpha^2 = \frac{N_2(N_1-4)(N_1-1)}{(N_2-1)(N_1-3)(N_1-2)} \quad (5.5)$$

식 (5.4)를 이용하여 확충된 자료의 평균과 분산은 다음 식 (5.6)과 같이 주어진다(Matalas와 Jacobs, 1964).

$$\hat{\mu}_y = \hat{\mu}_{y_1}+\frac{N_2}{N_1+N_2}\hat{\rho}_{xy}\frac{\hat{\rho}_{y_1}}{\hat{\rho}_{x_1}} \quad (5.6)$$

$$\hat{\sigma}_y^2 = \frac{1}{N_1+N_2-1}\Big[(N_1-1)\hat{\sigma}_{y_1}^2 + (N_2-1)\hat{\rho}_{xy}\frac{\hat{\rho}_{y_1}^2}{\hat{\rho}_{x_1}^2}\hat{\sigma}_{x_2}^2+\frac{N_1N_2}{N_1+N_2} \hat{\rho}_{xy}^2\frac{\hat{\rho}_{y_1}^2}{\hat{\rho}_{x_1}^2}(\hat{\mu}_{x_2}-\hat{\mu}_{x_1})^2+(N_2-1) \theta\alpha^2(1-\hat{\rho}_{xy}^2)\hat{\sigma}_{y_1}^2\Big] \quad (5.7)$$

또한 Vogel과 Stedinger(1985)는 보다 발전된 평균과 분산의 추정량을 유도하였으나 본 강좌에서는 생략하기로 한다.

이와 같이 자료기간이 짧은 지점의 자료를 확충하는 경우에 1) 긴 자료기간을 갖는 지점의 자료를 짧은 지점의 자료 확충에 사용할 수 있는가? 2) 확충된 자료가 대상 지점의 매개변수값의 정도를 높이는데 기여할 수 있는가? 와 같이 두 가지 의문이 제기될 수 있으며 이에 대한 기준이 필요하다.

첫번째 의문은 자료 X와 Y의 상관성을 의미하는 것으로서, 두 지점 자료간의 상관계수값이 0과 크게 다르면 자료 X는 자료 Y의 정보를 충분히 포함하고 있으므로 자료 Y를 이용하여 자료 X를 확충할 수 있다. 이 경우 한계상관계수값(critical value of correlation coefficient)은 자유도가 (N_1-2)이고 분위수(quantile)가 $1-\alpha/2$인 student t-분포값을 이용하여 다음 식과 같이 주어지므로

$$\rho_c = \frac{t_{1-\alpha/2}(N_1-2)}{\sqrt{(N_1-2)+t_{1-\alpha/2}^2(N_1-2)}} \quad (5.8)$$

다음 식을 만족하는 경우에 유의수준 α에서 자료 X와 Y가 0의 상관계수를 갖는다는 귀무가설은 기각된다.

$$|\hat{\rho}_{xy}| > \rho_c \quad (5.9)$$

이와 같이 자료 X와 Y의 상관계수가 0이 아닌 경우에는 자료 Y의 확충을 위하여 자료 X를 이용할 수 있는 것이다.

두 번째 의문은 주변 지점의 자료를 이용하여 확충된 자료가 확충되기 이전과 비교하여 신뢰할 수 있는가 하는 질문이다. 이런 경우에는 확충전 자료의 매개변수 추정량의 분산값과 확충후의 분산값을 다음 식과 같이 비교하여 검증할 수 있다.

$$\sigma^2(\hat{\theta}_e) < \sigma^2(\hat{\theta}_o) \quad (5.10)$$

여기서 $\sigma^2(\hat{\theta}_o)$, $\sigma^2(\hat{\theta}_e)$는 각각 확충전 및 확충후의 매개변수 추정량의 분산값으로 확충후의 분산값

이 확충전의 분산값보다 작으면 앞서 설명한 선형 회귀방정식을 이용하여 자료를 확충하는 것이 유익하다고 할 수 있는 것이다.

식 (5.10)은 다음과 같이 표현할 수 있으며

$$I = \sigma^2(\hat{\theta}_o)/\sigma^2(\hat{\theta}_e) \tag{5.11}$$

여기서 I값이 1보다 크면 자료확충으로 인한 매개변수 추정에 도움이 된다고 할 수 있다.

위에서 언급한 방법 외에도 매개변수 추정의 향상을 판단하는 방법으로 유효자료수(effective sample size) 또는 상당자료수(equivalent sample size)를 정의하여 구하는 방법이 있다. 유효자료수는 다음과 같이 정의된다.

$$N_e = N_1 I \tag{5.12}$$

여기서 I 값이 1보다 크면 유효자료수는 확충전의 자료수 N_1보다 크게 되므로 확충으로 인한 이득(gain)이 생기게 되며 이를 식으로 표시하면 다음과 같다.

$$N_g = N_e - N_1 \tag{5.13}$$

위에서 설명한 내용을 요약 정리하면 표 5.1과 같다.

표 5.1 자료확충에 의한 매개변수 추정 판단 기준

분 산	판 단 기 준			
	Index	Effective sample size	Gain of Information	선형회귀방정식 사용여부
$\sigma^2(\hat{\theta}_e) < \sigma^2(\hat{\theta}_o)$	$I > 1$	$N_e > N_1$	$N_g > 0$	Yes
$\sigma^2(\hat{\theta}_e) \geq \sigma^2(\hat{\theta}_o)$	$I \leq 1$	$N_e \leq N_1$	$N_g \leq 0$	No

자료의 평균과 분산의 예를 설명하면 다음과 같다. 평균을 개선하기 위한 유효자료수는 다음 식과 같이 주어지며(Langbein, 1960)

$$N_e = \frac{N_1 + N_2}{1 + N_2(1-\rho_{xy}^2)/(N_1-2)} \tag{5.14}$$

식 (5.13)에서 N_g값이 0보다 크기 위해서는 다음의 조건을 만족해야 한다.

$$\hat{\rho}_{xy} > \sqrt{2/N_1} \tag{5.15}$$

한편 평균에 대한 분산은 다음 식과 같이 주어지며(Fiering, 1963; Matalas와 Jacobs, 1964)

$$\sigma^2(\hat{\mu}_y) = \frac{\sigma^2{}_y}{N_1}\left[1 - \frac{N_2}{N_1+N_2}\left(\hat{\rho}_{xy}^2 - \frac{1-\hat{\rho}_{xy}^2}{N_1-3}\right)\right] \tag{5.16}$$

확충된 자료에 대한 평균값이 개선되기 위해서는 $\sigma^2(\hat{\mu}_y) < \sigma^2(\hat{\mu}_{y_1})$인 조건을 만족해야 하며 이를 위해서는 다음의 조건식을 만족해야 한다.

$$\hat{\rho}_{xy} > \sqrt{1/(N_1-2)} \tag{5.17}$$

식 (5.17)의 우변의 값을 자료 확충을 통하여 평균을 개선하기 위한 한계최소상관계수(critical minimum correlation coefficient) 라고 한다. 그러므로, 자료를 확충하여 구한 상관계수가 이 값보다 작으면 자료 확충이 도움이 되지 않는다고 할 수 있다.

마찬가지로 식 (5.4)를 이용하여 자료를 확충하는 경우, 확충된 자료의 분산값의 분산은 다음 식과 주어진다(Matalas와 Jacobs, 1964)

$$\sigma^2(\hat{\sigma}_y^2) = \frac{2\sigma_y^4}{N_1-1} + \frac{N_2\sigma_y^4}{(N_1+N_2-1)^2}(A\hat{\rho}_{xy} + B\hat{\rho}_{xy}^2 + C) \tag{5.18}$$

여기서

$$A = \left[\frac{(N_2+2)(N_1-6)(N_1-8)}{(N1-3)(N_1-5)} - \frac{8(N_1-4)}{(N_1-3)} - \frac{2N_2(N_1-4)^2}{(N_1-3)^2}\theta^2 + \frac{4(N_1-4)}{(N_1-3)}\theta^2 + \frac{N_1N_2(N_1-4)^2}{(N_1-3)^2(N_1-2)}\theta^4\right] \tag{5.19a}$$

$$B=\left[\frac{6(N_2+2)(N_1-6)}{(N_1-3)(N_1-5)}+\frac{2(N_1^2+N_1-14)}{(N_1-3)}\right.$$
$$-\frac{2(N_1+N_2-1)(N_1-4)}{(N_1-3)}(1-\theta^2)$$
$$+\frac{2N_2(N_1-4)(N_1-5)}{(N_1-3)^2}\theta^2$$
$$+\frac{2(N_1-4)(N_1+3)}{(N_1-3)}\theta^2$$
$$\left.-\frac{2N_1N_2(N_1-4)^2}{(N_1-3)^2(N_1-2)}\theta^4\right] \qquad (5.19b)$$

$$C=\left[\frac{2(N_1+1)}{N_1-3}+\frac{3(N_2+2)}{(N_1-3)(N_1-5)}\right.$$
$$-\frac{(N_1+1)(2N_1+N_2-2)}{(N_1-1)}$$
$$+\frac{2(N_1+N_2-1)(N_1-4)}{(N_1-3)}(1-\theta^2)$$
$$+\frac{2N^2(N_1-4)}{(N_1-3)^2}\theta^2$$
$$\left.+\frac{N_1N_2(N_1-4)^2}{(N_1-3)(N_1-2)}\theta^4\right] \qquad (5.19b)$$

여기서도 평균의 경우와 마찬가지로 확충된 자료에 대한 분산값이 개선되기 위해서는 $\sigma^2(\hat{\sigma}_y^2)<\sigma^2(\hat{\sigma}_{y_i}^2)$인 조건을 만족해야 하며 이를 위해서는 다음의 조건 식을 만족해야 한다.

$$|\hat{\rho}_{xy}| > \left(\frac{-B\pm\sqrt{B^2-4AC}}{2A}\right)^{1/2} \qquad (5.20)$$

식 (5.20)의 우변의 값을 자료 확충을 통하여 분산값을 개선하기 위한 한계최소상관계수라 한다.

식 (5.4)를 이용하여 자료를 확충하는 경우 오차항 ε_i는 무작위성을 가지므로, 난수(random number) 발생 시마다 상이한 결과를 갖게 된다. 이 경우 확충전 자료와 동일한 평균 및 분산값을 갖는 자료로 확충하기 위하여 Hirsch(1982), Hirsch와 Gilroy(1984), Vogel과 Stedinger (1985), Grygier 등(1989)이 새로운 모형을 제안하였으나 본 강좌에서는 참고문헌으로 대체하기로 한다.

5.1.2 다중회귀방정식

대상 지점의 자료의 확충은 2개 또는 그 이상의 지점 자료를 사용하는 다중회귀방정식을 이용하여 구할 수 있다. 대상 지점의 자료를 y, 주변 i 지점의 자료를 $x^{(i)}$ (i=1, . . ., m)라 하면 다음과 같이 나타낼 수 있다.

$y_1, y_2, \cdots, y_{N_1}$
$x_1^{(1)}, x_2^{(2)}, \cdots, x_{N_1}^{(1)}, x_{N_1+1}^{(1)}, x_{N_1+N_2}^{(1)}$
$x_1^{(m)}, x_2^{(m)}, \cdots, x_{N_1}^{(m)}, x_{N_1+1}^{(m)}, x_{N_1+N_2}^{(m)}$

여기서 동시 자료는 다변량정규분포(multivariate normal distribution)라고 가정한다. 위와 같이 주어진 자료를 이용하여 대상 지점의 자료를 확충하는 다중회귀방정식은 다음과 같다(Gilroy, 1970).

$$y_t=\hat{a}+\sum_{i=1}^{m}\hat{b}_i x_t^{(i)}+(1-\hat{R}^2)^{1/2}\alpha\theta\hat{\sigma}_y\varepsilon_t \qquad (5.21)$$

여기서 $\hat{R}$은 자료 y와 $x^{(i)}$ 사이의 다중상관계수(multiple correlation coefficient)이며 α는 다음 식과 같이 주어진다.

$$\alpha=\left[\frac{N_2(N_1-2m-2)(N_1-1)}{(N_1-1)(N_1-m-2)(N_1-m-1)}\right]^{1/2} \qquad (5.22)$$

Gilroy(1970)는 식 (5.21)을 이용하여 확충된 자료의 평균 및 분산추정량의 분산을 유도하였으나, 분산추정량의 분산식이 잘못되어 Moran (1974)이 분산에 대한 최우도추정량의 분산식을 새롭게 유도하였으며, 허준행(1994)은 이를 Matalas와 Jacobs가 유도한 형태로 변환시켜 확충된 자료의 평균 및 분산이 개선되기 위한 한계최소상관계수를 유도하고 주변 지점의 수에 따른 표와 그림을 제시하였으니 참고하길 바란다. 다중회귀방정식을 사용하는 경우에도 선형회귀방정식을 사용하는 경우와 마찬가지로, 확충된 자료의 평균 및 분산추정량의 분산값이 확충전 추정량의 분산값

보다 작아야 한다.

또한 다중회귀방정식을 사용하여 자료를 확충하는 경우 확충전 자료와 동일한 평균 및 분산값을 갖는 자료로 확충하는 과정은 앞서 언급한 참고문헌의 내용을 응용하여 구할 수 있으므로 본 강좌에서는 내용을 생략하기로 한다.

5.1.3 자기상관을 갖는 경우

앞에서 설명한 방법들은 대상 자료가 자기상관성이 없다(serially uncorrelated)는 가정 하에서 유도된 것으로 실제 자료인 경우 상관성이 있는 경우가 많다. 이런 경우에는 상당독립자료수(equivalent independent sample size)를 이용한 회귀방정식을 사용하거나(Barlett, 1935) 또는 변환함수모형(transfer function model)과 같은 시간종속적인 모형을 사용하는 방법이 있다(Beauchamp 등, 1989). 자세한 내용은 본 강좌에서는 생략하기로 한다.

5.2 연자료의 결측치 보완 및 자료 확충

5.2.1 단일확률변수인 경우

대상 지점의 연자료만을 이용하여 자료를 확충하는 경우, 자료간의 자기상관성이 존재하면 AR(1) 모형을 사용할 수 있다.

$$y_t=\hat{\mu}_y+\hat{\rho}_1(y_{t-1}-\hat{\mu}_y)+\hat{\sigma}_y\sqrt{1-\hat{\rho}_1^2}\,\varepsilon_t \qquad (5.23)$$

여기서 $\hat{\rho}_1$은 lag-1인 자기상관계수(serial correlation coefficient)이고 ε_t는 표준정규독립변수이다. 식 (5.23)을 사용할 때 오차항 ε_t는 대부분의 경우 사용하지 않으며, 자기상관계수의 값이 음수인 경우에는 주의가 필요하다. 또한 결측치가 연속적으로 긴 경우에는 AR(1)모형을 사용하지 않는 것이 일반적이다.

5.2.2 2개 이상의 확률변수인 경우

대상 지점 주변 2개 이상의 연자료를 사용하는 방법 중에서 정규비방법(normal ratio method), 가중평균법(weighted average method), 선형회귀방정식 등에 대하여 설명하기로 한다.

① 정규비방법

대상 지점의 자료를 $y_t(t=1, \ldots, N)$, 주변 m개 지점의 자료를 $x_t^{(i)}$ $(i=1, \ldots, m)$라고 할 때 다음의 관계식을 만족하면

$$0.9\hat{\mu}_y \le \hat{\mu}_x^{(i)} \le 1.1\hat{\mu}_y \qquad (5.24)$$

대상 지점 결측치는 다음 식을 이용하여 구할 수 있다(Linsley 등, 1982).

$$y_t = \frac{1}{m}\sum_{i=1}^{m} x_t^{(i)} \qquad (5.25)$$

여기서 $\hat{\mu}_y$, $\hat{\mu}_x^{(i)}$는 각각 결측치를 갖고 있는 대상지점 및 i지점의 평균값이다. 만약에 식 (5.24)의 조건을 만족하지 못하는 경우에는 다음 식을 이용하여 결측치를 보완할 수 있다.

$$y_t = \frac{\hat{\mu}_y}{m}\sum_{i=1}^{m}\frac{x_t^{(i)}}{\hat{\mu}_x^{(i)}} \qquad (5.26)$$

② 가중평균법

이 방법은 미국 기상청(NWS, 1972)에서 추천하는 방법으로 그림 5.1과 같이 대상 지점을 중심으로 동서남북 4개의 상한(quadrants)으로 구분한 뒤 각 상한별로 대상 지점에 제일 가까운 지점의 자료를 이용하여 결측치를 보완하는 방법으로 다음 식과 같이 주어진다.

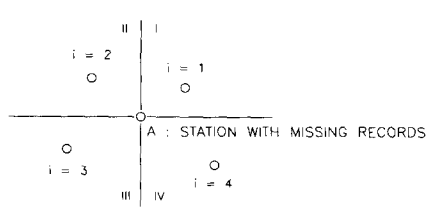

그림 5.1 가중평균법의 적용례(NWS, 1972)

$$y_t = \sum_{i=1}^{4} w_i x_t^{(i)} / \sum_{i=1}^{4} w_i \qquad (5.27)$$

여기서 $w_i = 1/L_i^2$ 이고, L_i는 대상 지점과 주변 i지점 사이의 거리다. 이 방법은 적어도 2개 이상의 주변 지점 자료를 이용하는 경우에만 사용된다.

③ 선형회귀방정식

앞에서 단순 및 다중회귀방정식을 이용하여 자료를 확충하는 방법에 대해서 설명한 바 있다. 여기서는 동일시간의 자료뿐 아니라 하나 앞선 시간의 자료를 이용하여 구하는 방법에 대하여 설명하기로 한다. 이와 같은 경우 다중회귀방정식은 다음 식과 같이 주어진다.

$$\begin{aligned} y_t = a &+ b_1 x_t^{(1)} + b_2 x_t^{(2)} + \cdots + b_m x_t^{(m)} + b_{m+1} x_{t-1}^{(1)} \\ &+ b_{m+2} x_{t-1}^{(2)} + \cdots + b_{2m} x_{t-1}^{(m)} + b_{2m+1} y_{t-1} \\ &+ \theta \alpha \sqrt{1-R^2}\, \sigma_y \varepsilon_t \end{aligned} \qquad (5.28)$$

식 (5.28)에서 보는바와 같이 독립변수의 수는 2m+1이며, 매개변수를 구하기 위하여 사용되는 자료수는 N−1이다. 또한 단순회귀방정식인 경우 (m=1) 식 (5.28)은 다음과 같이 쓸 수 있다.

$$\begin{aligned} y_t = a &+ b_1 x_t + b_2 x_{t-1} + b_3 y_{t-1} \\ &+ \theta \alpha \sqrt{1-R^2}\, \sigma_y \varepsilon_t \end{aligned} \qquad (5.29)$$

5.3 계절 자료의 결측치 보완 및 자료 확충

대상자료가 계절자료일 때 단일확률변수 및 다중확률변수인 경우에 대해 알아보기로 하자.

5.3.1 단일확률변수인 경우

① 정규비방법

대상 자료를 $y_{\nu,\tau}$ ($\nu=1, \ldots, N$ 자료수 : $\tau=1, \ldots, \omega$ 계절수)라 하면 각 계절별 평균은 다음 식과 같다.

$$\hat{\mu}_\tau = \frac{1}{N_\tau} \sum_{\nu=1}^{N_\tau} y_{\nu,\tau} \qquad (5.30)$$

여기서 N_τ는 계절별 자료수로서 결측치가 없는 경우에는 $N_\tau = N$이다.

결측치를 $y_{\nu',\tau'}$라고 하고 결측치를 포함하는 계절 τ'과 나머지 계절에 대해 정규비방법을 적용하면

$$\frac{y_{\nu',\tau'}}{\hat{\mu}_{\tau'}} = \frac{\sum_{\tau=1}^{\omega} y_{\nu',\tau'}}{\sum_{\tau=1}^{\omega} \mu_\tau} = \frac{y_{\nu',\tau'} + \sum_{\tau \neq \tau'}^{\omega} y_{\nu',\tau}}{\hat{\mu}} \qquad (5.31)$$

여기서 $\hat{\mu} = \sum_{\tau=1}^{\omega} \hat{\mu}_\tau$로 식 (5.31)을 $y_{\nu',\tau'}$에 대해서 풀면 다음과 같이 주어지며

$$y_{\nu',\tau'} = \frac{\hat{\mu}_{\tau'} + \sum_{\tau \neq \tau'} y_{\nu',\tau}}{\hat{\mu} - \hat{\mu}_{\tau'}} \qquad (5.32)$$

식 (5.32)를 이용하여 결측치를 보완할 수 있다. 만약에 2개 이상의 결측치가 있는 경우에는 선형방정식을 작성하여 풀어야 한다.

② PAR(1) 모형(Lag−1 Periodic Autoregressive Model)

대상 자료의 평균, 분산 및 상관계수가 주기성을 갖고 PAR(1)모형을 만족하는 경우 결측치를 보완하는 식은 다음과 같다.

$$\begin{aligned} y_{\nu,\tau} = \hat{\mu}_\tau &+ \hat{\rho}_{1,\tau} \frac{\hat{\sigma}_\tau}{\hat{\sigma}_{\tau-1}} (y_{\nu,\tau-1} - \hat{\mu}_{\tau-1}) \\ &+ \theta \hat{\sigma}_\tau \sqrt{1-\hat{\rho}^2_{1,\tau}}\, \varepsilon_{\nu,\tau} \end{aligned} \qquad (5.33)$$

여기서 $\hat{\rho}_{1,\tau}$는 계절 τ의 lag-1 주기상관계수이며, $\varepsilon_{\nu,\tau}$는 독립표준정규변수이다. 그러므로, 식 (5.33)에서 $y_{\nu,\tau-1}$을 알면 결측치 $y_{\nu,\tau}$를 구할 수 있다. 만약 $\tau-1$계절에 결측치가 있으면 식 (5.33)을 다음과 같이 쓸 수 있으며

$$\begin{aligned} y_{\nu,\tau+1} = \hat{\mu}_{\tau+1} &+ \hat{\rho}_{1,\tau+1} \frac{\hat{\sigma}_{\tau+1}}{\hat{\sigma}_\tau} (y_{\nu,\tau} - \hat{\mu}_\tau) \\ &+ \theta \hat{\sigma}_{\tau+1} \sqrt{1-\hat{\rho}^2_{1,\tau+1}}\, \varepsilon_{\nu,\tau+1} \end{aligned} \qquad (5.34)$$

이를 $y_{\nu,\tau}$에 대하여 다시 정리하면 다음 식과 같다.

$$y_{\nu,\tau} = \hat{\mu}_{\tau} + \frac{\hat{\sigma}_{\tau}}{\hat{\rho}_{1,\tau+1}\,\hat{\sigma}_{\tau+1}}(y_{\nu,\tau+1}-\hat{\mu}_{\tau+1}) + \theta\hat{\sigma}_{\tau}\sqrt{\frac{1}{\hat{\rho}^2_{1,\tau+1}}-1}\;\varepsilon_{\nu,\tau+1} \quad (5.35)$$

그러므로 $y_{\nu,\tau+1}$을 알면 $y_{\nu,\tau}$를 구할 수 있다. 그러나 $y_{\nu,\tau-1}$과 $y_{\nu,\tau+1}$ 값을 둘 다 모르면 결측치는 보완될 수 없다.

5.3.2 2개의 확률변수인 경우

대상 자료를 $y_{\nu,\tau}$ 다른 지점의 자료를 $x_{\nu,\tau}$($\nu=1, \ldots, N$ 자료수 ; $\tau=1, \ldots, \omega$ 계절수)라 할 때 두 자료간의 단순회귀방정식은 다음과 같이 표현된다.

$$y_{\nu,\tau} = \hat{a}_{\tau} + \hat{b}_{\tau}\,x_{\nu,\tau} + \hat{\alpha}\,\theta\,\sqrt{1-\hat{\rho}^2_{\tau}}\;\sigma_{\tau}(y)\;\varepsilon_{\nu,\tau} \quad (5.36)$$

여기서 $\hat{a}_{\tau}$, $\hat{b}_{\tau}$, $\hat{\rho}_{\tau}$는 계절 τ에 대한 계수 및 상관계수이며, $\hat{\sigma}_{\tau}(y)$는 $y_{\nu,\tau}$의 주기표준편차(periodic standard deviation), $\hat{\alpha}$는 식 (5.5)를 통하여 구해지는 상수이다.

5.3.3 다중 확률변수인 경우

① 정규비 방법

m개 지점의 자료가 있고, i지점의 ν연도 τ계절의 자료 $x^{(i)}_{\nu,\tau}$가 결측되었다고 하면 정규비 방법에 의한 결측값은 다음과 같이 주어진다(Linsley 등, 1982).

$$x^{(i)}_{\nu,\tau} = \frac{\mu_{\tau}^{(i)}}{\sum_{k\neq i}^{m}\phi_k}\sum_{k\neq i}^{m}\frac{\phi_k x^{(k)}_{\nu,\tau}}{\mu_{\tau}^{(k)}} \quad (5.37)$$

여기서 ϕ_k는 가중값으로 일반적으로 1을 사용한다. 이 경우 다음의 관계식을 만족하면

$$0.9\hat{\mu}^{(i)}_{\tau} \le \hat{\mu}^{(k)}_{\tau} \le 1.1\hat{\mu}^{(i)}_{\tau} \quad \text{for all } k\neq i \quad (5.38)$$

결측치 $x^{(i)}_{\nu,\tau}$는 다음 식으로부터 보완될 수 있다.

$$x^{(i)}_{\nu,\tau} = \frac{1}{\sum_{k\neq i}^{m}\phi_k}\sum_{k\neq i}^{m}\phi_k x^{(k)}_{\nu,\tau} \quad (5.39)$$

② 다중회귀방정식

결측치를 포함하는 대상 지점의 자료를 $y_{\nu,\tau}$, m개의 주변 자료를 $x^{(i)}_{\nu,\tau}$ (i=1, . . ., m)라고 하면 다중회귀방정식은 다음과 같이 쓸 수 있다.

$$y_{\nu,\tau} = a_{\tau} + b_{1,\tau}x^{(1)}_{\nu,\tau} + b_{2,\tau}x^{(2)}_{\nu,\tau} + \cdots + b_{m,\tau}x^{(m)}_{\nu,\tau} + \alpha_{\tau}\theta\sqrt{1-R^2_{\tau}}\;\sigma_{\tau}(y)\varepsilon_{\nu,\tau} \quad (5.40)$$

여기서 $a_{\tau}+b_{1,\tau}, \ldots, b_{m,\tau}$는 계절 τ에 대한 회귀계수, R_{τ}는 자료 $y_{\nu,\tau}$와 $x^{(i)}_{\nu,\tau}$ 사이의 다중상관계수(multiple correlation coefficient), $\sigma_{\tau}(y)$는 $y_{\nu,\tau}$의 주기표준편차, α는 식 (5.22)로부터 정해지는 상수이다.

다중회귀방정식을 이용하여 결측치를 보완하는 경우에 결측치를 갖는 계절 τ보다 하나 앞선 계절 $(\tau-1)$ 자료 또는 하나 뒤의 계절$(\tau+1)$ 자료, 또는 앞뒤 계절 모두를 포함하는 방정식을 사용할 수 있으며, 이경우의 다중회귀방정식은 다음 식과 같이 주어진다.

$$y_{\nu,\tau} = a_{\tau} + b_{1,\tau}y_{\nu,\tau-1} + b_{2,\tau}y_{\nu,\tau+1} + b_{3,\tau}x^{(1)}_{\nu,\tau} + \cdots + b_{m+2,\tau}x^{(m)}_{\nu,\tau} + b_{m+3,\tau}x^{(1)}_{\nu,\tau-1} + \cdots + b_{2m+2,\tau}x^{(m)}_{\nu,\tau-1} + b_{2m+3,\tau}x^{(1)}_{\nu,\tau+1} + \cdots + b_{3m+2,\tau}x^{(m)}_{\nu,\tau+1} + \alpha_{\tau}\theta\sqrt{1-R^2}\sigma_{\tau}(y)\varepsilon_{\nu,\tau} \quad (5.41)$$

다중회귀방정식을 이용하여 결측치를 보완하는 방법으로 자료를 표준화시켜 연자료의 경우와 같이 적용하는 방법이 있다. 결측치를 포함하는 계절자료는 다음과 같이 표준화시킬 수 있다.

$$u_t = \frac{y_{\nu,\tau}-\hat{\mu}_{\tau}(y)}{\hat{\sigma}_{\tau}(y)} \quad \text{for } t=(\nu-1)\omega+\tau \quad (5.42)$$

여기서 $\hat{\mu}_{\tau}(y)$, $\hat{\sigma}_{\tau}(y)$는 각각 계절 τ에 대한 주기평균 및 표준편차이다. 마찬가지로 주변 지점 자

료도 다음과 같이 표준화시킬 수 있다.

$$z_t^{(i)} = \frac{y_{\nu,\tau}^{(i)} - \hat{\mu}_\tau^{(i)}(x)}{\hat{\sigma}_\tau^{(i)}(x)} \tag{5.43}$$

여기서, $\mu_\tau^{(i)}(x), \sigma_\tau^{(i)}(x)$는 각각 i 지점의 계절 τ에 대한 주기평균 및 표준편차이다.

결측치를 포함한 지점의 표준화된 자료 u_t와 주변지점의 표준화된 자료 $z_t^{(i)}$을 이용한 선형회귀방정식은 다음과 같이 쓸 수 있다.

$$u_t = a + b_1 z_t^{(1)} + b_2 z_t^{(2)} + \cdots + b_m z_t^{(m)} + \theta\alpha\sqrt{1-R^2}\,\sigma_u \varepsilon_t \tag{5.44}$$

여기서 a, b_1, . . ., b_m은 회귀계수이며, R은 u_t와 $z_t^{(i)}$사이의 다중상관계수, α는 식 (5.22)로부터 구해지는 상수, σ_u는 자료 u_t의 표준편차, ε_t는 오차를 나타내는 표준정규변수이다.

5.4 맺음말

아래와 같이 "수문통계학의 기초"란 제목하에 지난 5번의 기술강좌를 통하여 수문 또는 수자원 자료를 통계학적으로 처리하는데 있어서 기본적으로 알아야 하는 용어 및 방법에 대하여 간단하게 설명하였다.

1. 기초 통계학(Basic Statistics)
2. 빈도해석(비매개변수적·매개변수적 방법) (Nonparametric and Parametric Frequency Analysis)
3. 검정방법(Various Tests)
4. 자료의 경향 및 변동 측정방법(Detection of Changes and Trend in Data)
5. 결측치보완 및 자료확충방법(Filling in Missing Data and Extension of Records)

본 강좌 및 참고문헌을 통하여 수문 및 수자원 자료를 통계학적으로 처리하는 실무자들에게 도움이 될 수 있기를 바라며 또한 본 강좌를 통하여 통계학적 처리 방법에 발전이 있기를 바랍니다. 마지막으로 그 동안 많은 격려와 조언을 해주신 여러분에게 감사를 드립니다.

참 고 문 헌

허준행 (1994). "수문자료의 확충을 위한 다중상관계수의 한계최소치의 유도." 한국수문학회지 제26권 제5호, pp. 133-140.

Barlett, M. S. (1935). "Some aspects of the time-correlation problem in regard to test of significance." *J. Royal Statis. Soc.*, Vol. 98, pp. 536-543.

Beauchamp, J. J., Downing, D. J., and Railsback, S. F. (1989). "Comparison of regression and time series methods for synthesizing missing streamflow records." *Water Resour. Bulletin*, Vol. 25, No. 5, pp. 961-975.

Fiering, M. B. (1963). Use of correlation to improve estimates of the mean and variance, *U.S. Geological Survey Prof. Paper* 434-C.

Gilroy, E. J. (1970). "Reliability of a variance estimate obtained from a sample augmented by multivariate regression." *Water Resour. Res.*, Vol. 6, No. 6, pp. 1595-1600.

Grygier, J. C., Stedinger, J. R., and Yin, H. B. (1989). "A generalized maintenance of variance extension procedure for extending correlated series." *Water Resour. Res.*, Vol. 25, No. 3, pp. 345-349.

Hirsch, R. M. (1982). "A comparison of four record extension techniques." *Water Resour. Res.*, Vol. 18, No. 4, pp. 1081-1088.

Hirsch, R. M. and Gilroy, E. J. (1984) "Methods of fitting a straight line to data. : examples in water resources." *Water Resour. Bulletin*, Vol. 20, No. 5, pp. 705-711.

Langbein, W. B. (1960). Hydrologic data networks and methods of extrapolating or extending avaliable hydrologic data. *Trans. Interregional Seminar on Hydrologic Networks and Methods*, Bangkok.

Linsley, R. K., Kohler, M. A., and Paulhus, J. L. H. (1982). *Hydrology for Engineers*. McGraw Hill Book Co., New York.

Matalas, N. C. and Jacobs, B. (1964). A correla-

tion procedure for augmenting hydrologic data, *U.S. Geological Survey Prof. Paper* 434-E.

Moran, M. A. (1974). "On estimators obtained from a sample augmented by multiple regression." *Water Resour. Res.*, Vol. 10, No. 1, pp. 81-85.

National Weather Service (1972). National weather service river forecast system procedures. *NOAA Tech. Memo.* NWS HYDRO-14, U.S. Dept. of Commerce, Silver Spring, MD.

Vogel, R. M. and Stedinger, J. R. (1985) "Minimum variance streamflow record augmentation procedures." *Water Resour. Res.*, Vol. 21, No. 5, pp. 715-723.

정 오 표

강 좌	위 치	수정전·후	내 용
II	식 (2.1) 밑	수 정 전	$[X_i'-\triangle X,\ X_i'+\triangle X]$
		수 정	$[X_i'-\triangle X/2,\ X_i'+\triangle X/2]$
III	식 (3.6)	수 정 전	$\left[\frac{-1-u_{1-\alpha/2}\sqrt{N-k-1}}{N-k},\ \frac{1-u_{1+\alpha/2}\sqrt{N-k-1}}{N-k}\right]$
		수 정	$\left[\frac{-1-u_{1-\alpha/2}\sqrt{N-k-1}}{N-k},\ \frac{-1+u_{1+\alpha/2}\sqrt{N-k-1}}{N-k}\right]$
	표 2. 표본상관계수	수 정 전	하한계와 상한계
		수 정	하한계와 상한계 값 수정(아래의 표 2 참고)
	식 (3.28) 밑	수 정 전	$d_{1-\alpha/2}(N)$
		수 정	$d_{\beta}(N)$
	표 7	수 정 전	β = 0.8 0.9 0.95 0.99
		수 정	β = 0.9 0.95 0.975 0.995
	식 (3.30) 밑	수 정 전	$W > w_{1-\alpha}(N)$
		수 정	$W \leq w_{1-\alpha}(N)$
	예제 3.11	수 정 전	$W > w_{1-\alpha}(N)$
		수 정	$W < w_{1-\alpha}(N)$

표 2. 표본상관계수

Lag	하 한 계	r_k	상 한 계
1	-.335	-.113	.284
2	-.340	-.076	.287
3	-.345	.212	.291
4	-.350	.094	.294
5	-.355	-.136	.298
6	-.361	.311	.302
7	-.366	.053	.306
8	-.372	-.053	.310
9	-.379	.029	.314
10	-.385	.030	.318