비모수적 기법에 의한 확률론적 저수지 유입량 예측

Probabilistic Reservoir Inflow Forecast Using Nonparametric Methods

이한구*, 김선기**, 조영현***, 정구열****

Han-Goo Lee, Sun-Gi Kim, Yong-Hyon Cho, Koo Yol Chong

요 지

추계학적 시계열 분석은 크게 수문자료의 장기간 합성과 실시간 예측으로 구분해 볼 수 있다. 장기간 합성은 주로 수문자료의 추계적 특성을 반영한 수자원 시스템의 운영율 개발에 이용되어 왔다. 반면에 실시간 예측은 수자원 시스템의 순응적(adaptive) 관리에 적용되고 있다. 두 개념의 차이로 전자는 시계열 자료를 합성하여 발생 가능한 모든 수문조합을 얻고자 하는 것이라면 후자는 전 시간의 수문량을 조건으로 하는 다음 시간의 값을 순응적으로 예측하는 것이라 할 수 있다. 수문자료의 합성과 예측에는 크게 결정론적, 확률론적 방법의 두 가지 대별될 수 있다. 결정론적 모델링 방법에는 인공신경망이나 Fuzzy 기법 등을 이용할 수 있으며, 확률론적 방법에는 ARMAX 등의 모수적 기법과 k-NN(k-nearest neighbor bootstrap resampling), KDE(kernel density estimates), 추계학적 인공신경망 등의 비모수적 기법으로 분류할 수 있다.

본 연구에서는 대표적 비모수적 기법인 k-NN를 이용하여 충주댐을 대상으로 월 및 일 유입량 자료의 예측 정도를 살펴보았다. 전 시간 관측치를 조건으로 하는 다음 시간의 조건부 확률분포를 구하여 평균값을 계산한 후 관측치와 비교함으로써 모형의 정도를 살펴보았다. 그리고 실시간 저수지 운영에 이 기법의 활용성과 장단점도 살펴보았다. 모형개발 절차로 모형의 보정을 거쳐 검증을 실시하였다. 결론적으로 월 및 일 유입량 예측에 k-NN 기법이 실무적으로 적용될 수 있었으며, 장점으로는 k-NN 기법이 다른 기법보다 모델링 절차가 비교적 쉬워 저수지 운영 최적화 등 타 시스템과의 연계에 수월함이 인식되었다.

핵심용어 : 유량예측, 비모수적 기법, 추계학적 모델링, k-NN

1. 서론

무작위 변수의 시계열은 결합 확률분포로 완벽하게 표현될 수 있으며, 이것은 조건부 확률분포와 주변 확률분포로 인수분해 된다. $X_t=f(X_{t-1},..,X_{t-p})+\varepsilon$ 형태의 함수관계를 해석하는 추계학적 모델링 과정은 바로 전 시간의 주어진 조건하에서 다음 시간의 상태변수에 대한 조건부 확률분포를 구하는 과정이라 할 때, 결합 확률은 실질적으로는 관심 대상은 아니라 할 수 있다. 여기서 ε은 오차를 의미한다. 조건부 확률의 종속구조를 간단히 하기 위해서 보통 유한의 과거 시간에 종속되는 Markov 과정을 이용하여 왔으며, 무작위 변수 X의 시계열을 $\{x_1, x_2, \cdots, x_t\}$라 표현할 때, 시

* 정회원 · 한국수자원공사 물관리센터 · E-mail : hglee@kwater.or.kr

** 회원 · 한국수자원공사 물관리센터 · E-mail : sgkim@kwater.or.kr

*** 회원 · 한국수자원공사 물관리센터 · E-mail : yhcho@kwater.or.kr

**** 정회원 · 한국수자원공사 물관리센터 · E-mail : kychong@kwater..or..kr

간 t의 상태 x_t는 식 (1)과 같이 조건부 확률을 이용하여 p 차수의 Markov 과정으로 모형화 할 수 있다.

$$P(x_t \mid x_{t-1},.....,x_{t-p}) = \frac{P(x_t, x_{t-1},.....,x_{t-p})}{P(x_{t-1},.....,x_{t-p})} = \frac{P(x_t, x_{t-1},.....,x_{t-p})}{\int P(x_t, x_{t-1},.....,x_{t-p})\, dx_t} \quad (1)$$

식 (1)의 해석적 해를 구함에 있어 차수 p가 높을 경우 분자의 결합 확률을 계산하여야 하나 문제는 이 과정이 매우 어렵다는 사실이다. 수문학자들은 이 문제를 해결하기 위해 크게 모수적 방법과 비모수적 방법으로 접근을 시도하였다. 모수적 방법으로는 고전적인 ARMA (auto regressive and moving average) 기법이 대표적이며, k-NN(k-nearest neighbor bootstrap resampling), KDE(kernel density estimates) 등의 비모수적 기법이 연구되어 왔다. 모수적 방법의 특징은 조건부 확률의 Markov 종속구조를 매개변수 Θ와 함께 식 (2)의 형태로 함수화 하고 보통 선형으로 간략화 하였으며, x_t의 확률분포를 얻기 위해서 오차 ε의 확률분포를 정규분포로 가정하여 왔다(Salas, 1993). 또한 Bayesian 기법과는 달리 매개변수를 확정론적으로 처리하며, 선형화의 단점을 보완하기 위해 인공신경망 기법도 연구된 바 있다.

$$X_t = f(X_{t-1},..,X_{t-p},\Theta) + \varepsilon \quad (2)$$

반면 비모수적 방법은 관측치에 식 (2)의 독립변수와 종속변수들 간의 모든 함수관계가 포함되어 있다고 판단함으로써 x_t의 조건부 확률을 함수에 의존하지 않고 관측치로부터 직접 구하는 방식이다. 따라서 이 방법은 data-driven 모델링 기법의 하나라 할 수 있다. 이 방법의 특징은 무작의 변수의 확률분포와 식 (2)와 같은 함수에 대한 사전정보를 완전히 배제하자는 것이다. 그러나 이 방법의 단점으로는 모수적 방법에 비해 x_t 확률분포의 신뢰도 구간이 넓으며, 관측치내에서 추정치를 구하는 관계로 외삽형식의 모델링은 불가능하다는 것이다. 즉 관측치에 발생 가능한 상태변수들의 종속관계를 충분히 담고 있지 못할 경우 모델링의 한계가 존재하게 된다(Sharma 등, 1997). 대표적인 비모수적 방법으로는 앞에서 언급한 k-NN, KDE 등이 존재하며, 추계학적 인공신경망 기법은 모수적 방법과 비모수적 방법의 중간에 있다고 할 수 있다. 비모수적 방법은 시계열 분석 외에 빈도해석, 공간분석 등에도 사용되고 있다.

본 연구에서는 대표적 비모수적 기법인 k-NN를 이용하여 충주댐을 대상으로 월 및 일 유입량 자료의 예측 정도를 살펴보았다. 전 시간 관측치를 조건으로 하는 다음 시간의 조건부 확률분포를 구하여 평균값을 계산한 후 관측치와 비교함으로써 모형의 정도를 살펴보았다. 그리고 실시간 저수지 운영에 이 기법의 활용성과 장단점도 살펴보았다. 모형개발 절차로 모형의 보정을 거쳐 검증을 실시하였다.

1.1 k-NN 기법

k-NN 방법의 기본 개념은 조건부 확률을 관측자료로부터 직접 산정하는 것이다. 일반적인 k-NN density estimator는 식 (1) (Silverman, 1986)로 표현되며, 이는 핵함수 (kernel function) $K(.)$와 주변 관측치 (neighbors)의 개수인 k로 이루어져 있다. k는 조건부확률의 smoothing factor이며, $K(.)$는 이 확률분포 양 끝단의 모양을 결정한다. 여기서, x는 확률변수의 임의 값이며, x_i는 i번째 관측값을, n은 관측값의 전체 개수를, r은 x와 x_i사이의 Euclidean 거리를, d는 상태공간의 차수를 의미한다.

$$f_{GNN}(x) = \frac{1}{r_k^d(x) n} \sum_{i=1}^{n} K\left(\frac{x - x_i}{r_k(x)}\right) \quad (3)$$

이를 구현하기 위해서는, 제일 먼저 현재시간에서의 상태벡터 $\{x_t, x_{t-1}, ..., x_{t-d}\}$와 유사한 주변 벡터(neighbor)들을 관측치로부터 알아내야 한다. 주변 벡터들을 결정할 때는 현재 상태벡터와 모든 과거 상태벡터와의 Euclidean 거리를 다음 식으로 구한 후, 가까운 순으로 k개를 선택하고, 이를 k Nearest Neighbors라 칭한다.

$$r_{ij} = \left(\sum_{m=1}^{d} w_m (x_{im} - x_{jm})^2 \right)^{1/2} \tag{4}$$

여기서, w는 가중치를 의미하며, x_i는 현재 상태벡터를, x_j는 j번째 과거 상태벡터를 의미한다. k를 정하는 방법은 다수 존재하나, k는 결과에 크게 영향을 미치지 않는 관계로 보통 $k=n^{1/2}$의 값으로 정하며, 자세한 내용은 Silverman (1986) 및 Lall(1996) 등에 자세히 소개되어 있다. 일단, k개의 주변 값들이 주어지면, 각각의 값들은 거리의 함수로 다음 식의 핵함수에 의해 일종의 확률 값을 지니게 된다. 즉, 거리가 가까울수록 확률은 커지며 멀수록 작아진다. 핵함수는 일종의 경험적 확률분포함수라 할 수 있다.

$$K(i,j) = \frac{1/j}{\sum_{j=1}^{k} 1/j} \tag{5}$$

핵함수가 결정되면, Bootstrap sampling 기법을 이용해서 k개의 주변 값들로부터 다수의 sample을 추출하고, 이로부터 다음 시간에서의 상태에 대한 통계적 추론이 가능해진다. 이 방법은 컴퓨터 프로그래밍으로 비교적 간단히 처리할 수 있다. 비록 매개변수 k가 존재하나, 이의 민감도가 낮은 점과, 또한 특정 확률분포를 가정하지 않고 경험적 확률분포를 고려한다는 점에서 대표적인 비모수적 방법으로 알려져 있다.

2. 월유입량 예측

본 연구에서는 월유입량 예측모형의 구조를 1차 지체(lag)의 자기회귀형으로 한정하였다. 물론 지체차수도 일종의 매개변수라 할 수 있는데 월단위 예측에는 보통 lag 1이 충분하다고 판단하였다. 1917년부터 2000년까지 84년간 충주댐 월유입량 자료를 이용하여 모형을 보정하였고, 2001부터 2005년의 자료를 이용하여 모형을 검증하였다. 식 (5) 핵함수의 매개변수인 k는 경험적으로 사용되는 $k=n^{1/2}$을 이용하거나 Akaike information criteria(AIC)와 Silverman(1986)이 제시하는 편차와 분산의 tradeoff 분석으로부터 결정되어 질 수 있다. 보다 실용적인 방법으로는 $k=n^{1/2}$ 주변의 값을 변화시켜 시행착오법으로 구하는 것이 가장 보편적일 것이다. 본 연구에서는 k의 민감도가 작은 관계로 경험적인 방법을 적용하였고, 84개의 월 자료가 존재하므로 n은 84이며 k는 약 9가 된다.

그림 1은 k=9인 모형의 보정결과를 그림 2는 검증결과를 각각 보여주고 있다. 그림 1에서는 복잡성을 피하기 위해 1995년부터 2000년까지만 도시하였고, 그 기간 동안 평균오차는 약 44㎥/초였다. 모형의 검증과정에서의 평균오차는 약 106㎥/초로 보정보다 약 2.5배의 오차를 보였다. 식 (5)의 이산형 핵함수에 의한 조건부 확률로부터 관측치를 선택하는 관계로 KDE 등의 다른 방법에 비해 거친 확률분포를 보이는 것으로 알려져 있다. 이 방법의 가장 큰 특징은 고도의 복잡한 계산 과정이 필요가 없어 수행시간이 빠르고 모형의 우수성도 뛰어나 실무적으로 활용성이 매우 높으며, 높은 차수의 Markov 모델에도 쉽게 적용이 가능하다. 따라서 수자원 분야의 의사결정지원시스템에 하나의 단위 모듈로 쉽게 연계될 수 있어 시스템의 통합화가 매우 수월하다. 모형의 검보정 결과로부터 무작위 특성이 심한 홍수와 가뭄시의 유량예측에는 한계를 보여주고 있음을 알 수 있다.

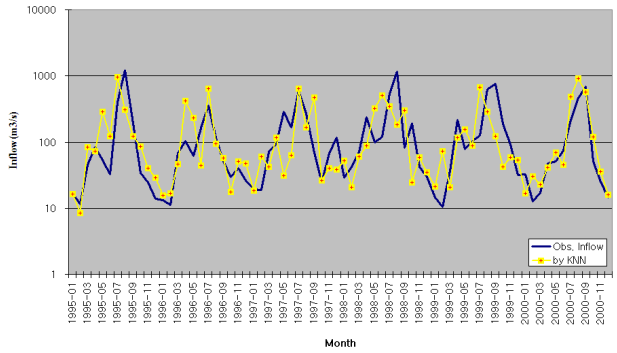

그림 1. 예측모형의 보정결과 (1995~2000)

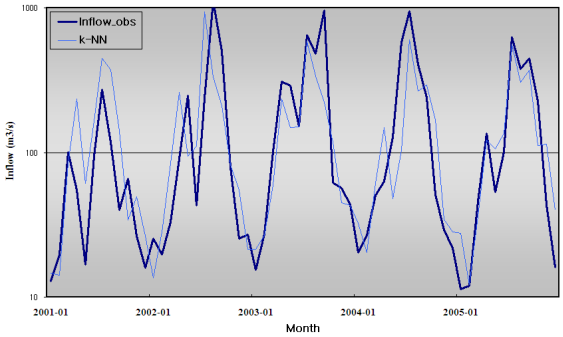

그림 2. 예측모형의 검증결과 (2001~2005)

3. 일유입량 예측

일단위 시계열 모형을 어떻게 적용하는 가에는 많은 방법이 존재한다. 예를 들면, 자기회귀의 차수를 무엇으로 할 것인가, 아니면 자기회귀이외의 다른 영향인자를 고려할 것인가, 고려한다면 그 변수의 차수는 무엇으로 할 것인가와 같은 선택사항에 따라 여러 형태의 시계열 모형을 구성될 수 있다. 본 연구에서는 다음 식과 같이 충주댐 일유입량에 대한 자기회귀의 차수를 1~6일의 범위 내에서, 유입량 외에 일강우량 (exogenous input variable)을 입력변수로 적용하여 차수를 1~10일의 범위 내에서 적용하여 보았다. 일종의 Auto-regressive Exogenous <ARX(p, q)> 모형으로 일유입량의 자기회귀 모형에 일강우량의 외부변수를 도입하였다. 여기서 p는 유입량, q는 강우량의 lag를 각각 의미한다. 일강우량을 도입한 이유는 일유입량의 시간적 관성만의 해석으로는 무작위 변동을 완벽하게 해석하기 어렵고, 예측대상일 이전에 발생한 강우량과 결합되어 해석될 때 무작위 변화에 더욱 완전하게 반응할 수 있기 때문이다. 유입량 차수 p를 6으로 한정한 이유는 관측 일유입량에 대한 Autocorrelation을 분석했을 때 수문학적으로 약 6~10일 정도의 시간적 관성을 보여 lag 1부터 최소치인 6까지 시간적 종속성이 유지되고 있는 것으로 판단하였다. 물론, lag의 한계를 결정하기 위한 절대적인 기준은 존재하지 않는다. 따라서 유입량과 강우량에 대한 lag를 변화시켜가면서 RMSE (root mean square error) 등의 통계치를 이용하여 적절한 모형을 선택하여야 한다. 강우량의 시간지체 차수인 q는 최대 10일까지 한정하였다. 적정 모형의 선택을 위해 정상 수문해인 1989년의 일유입량과 일강우량의 lag를 다르게 하면서 여러 조합을 시도하여 보았으나, 유입량 및 강우량 모두 lag-1의 ARX(1,1) 모형이 가장 적합하였다(그림 3). k는 위에서 살펴본 바와 같이 전체 관측자료 개수의 0.5승 ($k=n^{0.5}$)을 적용하였다.

그림 3에서 자기회귀 항목이 없이 강우량만으로 구성된 모형은 분명 적용하기에 무리가 있으며, 자기회귀 항목만에 의한 모형보다는 강우량이 포함된 경우가 더 나은 결과를 보임을 알 수 있다. 비록 모형의 보정결과(그림 4)로부터 관측치와 모의치가 잘 일치하고 있음을 알 수 있으나 월모형의 경우와 마찬가지로 그림 5의 검증과정에서 가뭄과 홍수의 극값들은 중간 유량보다 오차가 많음을 알 수 있다.

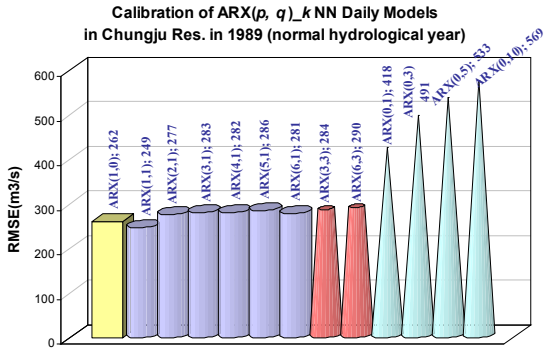

그림 3. 일모형 선정

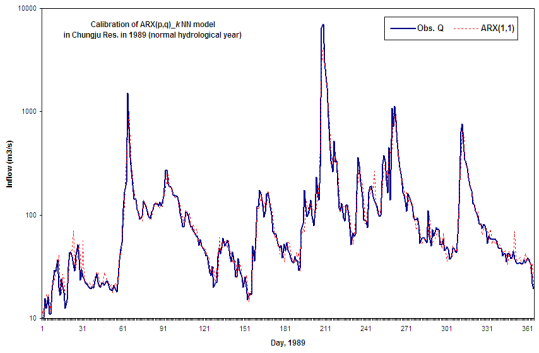

그림 4. 일모형의 보정 (1989)

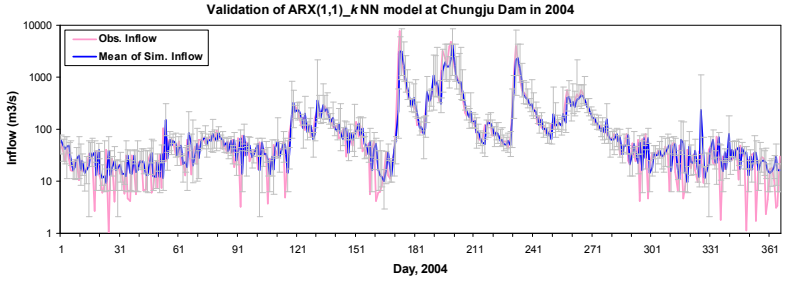

그림 5. 일모형 검증 (2004)

4. 결론

본 연구에서는 비모수적 기법인 k-NN방법을 사용하여 충주댐 유입량의 확률론적 예측모형을 구축하였다. 이 방법의 장점은 컴퓨터 프로그램화가 매우 쉬워 실무적 활용성이 뛰어나다는 것이며, 단점으로는 다른 모형과 마찬가지로 극값인 갈수량과 홍수량의 예측에 한계가 있다는 것이다. 저수지 운영에 있어 가장 중요한 사항은 수문예측의 불확실성에 의한 위험도를 의사결정에 반영하고 이러한 위험도가 연속적으로 관측되는 수문량을 반영하여 새롭게 구해짐으로써 순응적인 운영이 되도록 시스템의 통합화가 필요하다. 이러한 면에서 k-NN 기법은 다른 어느 모형보다 매우 우수하다고 말 할 수 있으며 저수지 운영 최적화 모형과 연계되어 실무적으로 활용이 충분히 가능하다.

참 고 문 헌

1. Lall, U., Sharma, Ashish (1996), “A nearest neighbor bootstrap for resampling hydrologic time series.” Water Res. Res.,32(3),679-693.
2. Salas, J. D. (1993) Analysis and modeling of hydrologic time series, in Handbook of Hydrology, edited by D.R. Maidment, McGraw-Hill, NewYork.
3. Sharma, A., Tarboton, D. G., Lall, U. (1997). “Stream flow simulation: A nonparametric approach.” Water Res. Res., 33(2), 291-308.
4. Silverman, B. W. (1986). Density Estimation for Statistics and Data Analysis, Chapmanand Hall, New York.