Journal of Korean Society of Water and Wastewater
Vol.23, No.2, pp.241-241, April, 2009

상하수도학회지, 논문
23권 2호, pp.241-249, 4월, 2009

물수요의 추세 변화의 적응을 위한 모델링 절차 제시:베이지안 매개변수 산정법 적용

Modeling Procedure to Adapt to Change of Trend of Water Demand: Application of Bayesian Parameter Estimation

이상은[1*] · 박희경[2]

Lee, Sangeun[1*] · PARK, Heekyung[2]

1 KAIST 응용과학연구소, 2 KAIST 건설및환경공학과

(2009년 1월 5일 접수; 2009년 4월 2일 수정; 2009년 4월 8일 채택)

Abstract

It is well known that the trend of water demand in large-size water supply systems has been suddenly changed, and many expansions of water supply facilities become unnecessary. To be cost-effective, thus, politicians as well as many professionals lay stress on the adaptive management of water supply facilities. Failure in adapting to the new trend of demand is sure to be the most critical reason of unnecessary expansions. Hence, we try to develop the model and modeling procedure that do not depend on the old data of demand, and provide engineers with the fast learning process. To forecast water demand of Seoul, the Bayesian parameter estimation was applied, which is a representative method for statistical pattern recognition. It results that we can get a useful time-series model after observing water demand during 6 years, although trend of water demand were suddenly changed.

Key words : **overcapacity, water treatment plants, trend of water demand, Bayesian Parameter Estimation**

주제어 : **과다용량, 정수장, 물수요 추세, 베이지안 매개변수 산정법**

1. 연구 필요성

최근 공공부문 투자의 효율성, 특히 정수장의 과다시설용량 문제를 다루기 위한 많은 논의가 진행되고 있다 (감사원, 2005; 박희경 등, 2007; 국회예산정책처, 2008). 특히, 박희경 등 (2007)에서는 전국 모든 지자체들의 정수장 용량의 최근 추이를 분석한 결과 시설용량의 적정성에 큰 문제가 있음을 지적하고 있다. 즉, 시·군단위의 소도시들의 경우 필요한 투자를 하지 못해서 정수장 용량이 부족한 실정임에 반해, 특광역시 등의 대도시들은 필요 이상으로 정수장 용량을 늘려왔던 것이다. 또한 이상은 등 (2009)에서는 **Fig. 1**과 같이 9개 대도시들의 예비용량이 서울시 천만 이상의 인구에 대한 일최대 물수요의 1.2배에 달하며, 계속적으로 가동률이 저하되고 있음을 보이고 있다.

전통적으로 상수도 시설의 용량계획은 상당히 먼 미래에 대한 수요를 예측한 뒤 이를 절대적으로 충당할 수 있는 시설을 대규모로 건설하는 예비용량가설 (Lauria, 1983) 또는 확장주의전략 (Ritzman and Krajewski, 2003)을 따른다. Gleick (2003)은 과다용량증설을 하게 되는 기술적인 원인으로 용량 산정시 물수요의 추세가 변함에도 불구하

* Corresponding author Tel:+, Fax:+, E-mail: eunism@kaist.ac.kr(Lee, S.)

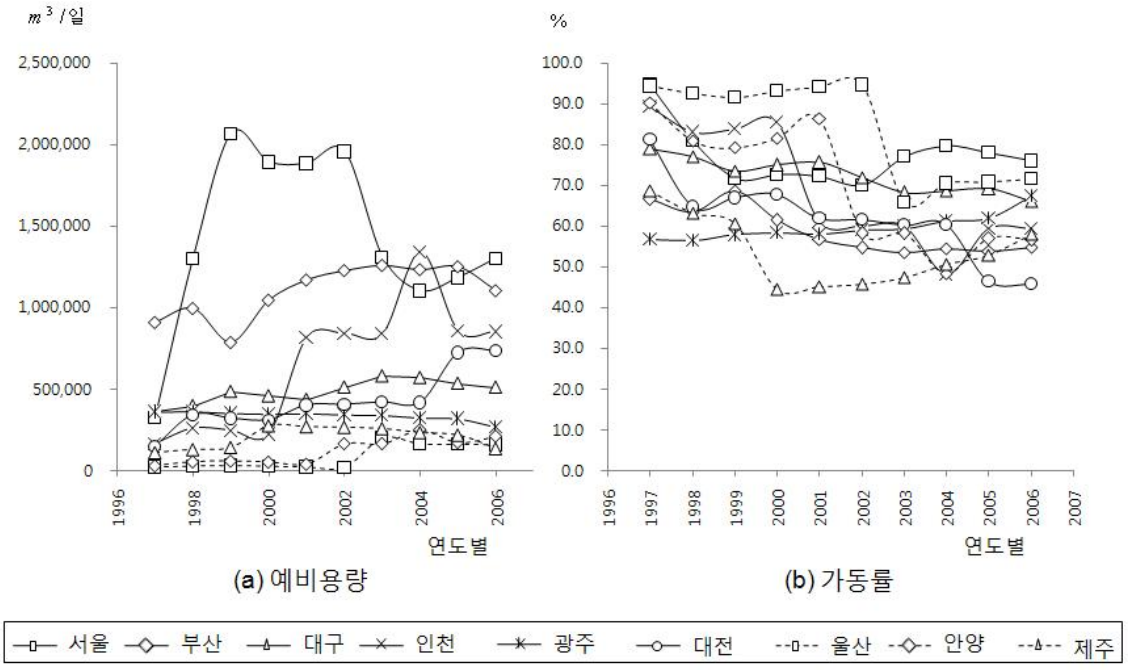


Fig. 1. 대표적인 대규모 지자체들의 정수장 가동 현황

고 기존의 물수요 예측 모형이 변한 추세를 잘 반영하지 못하는 점을 지적하고 있다. 이상은 등(2009)은 보다 구체적으로, 모형들은 장기 물수요 예측시 쉽고 자료요구가 적은 장점이 있지만 도시 인구가 더 이상 증가하지 않을 경우 경제적으로 위험한 예측 결과를 만듦을 설명하고 있다. 동 연구에서, 서울시 물수요에 대한 예를 들어, 만일 1997년에 정수장 용량 계획을 위해 1인당원단위모형 적용시 물수요 예측 결과는 **Fig. 2**와 같았다. 여기서, 오랜기간 동안 급수인구와 물수요간의 상관관계가 매우 높기 때문에 1995년부터 최근 3년 동안 물수요가 새로운 추세를 보인다는 사실을 예측 모형에 반영하기 힘든 점을 설명하고 있다. 그 결과, 1997년 이전인 검증구간에서는 결정계수 0.95 이상으로 성공적이었지만, 2006년 기준으로 예측 오차가 1.1백만 m^3/일에 달해 장래 물수요 예측을 완전히 실패하게 된다.

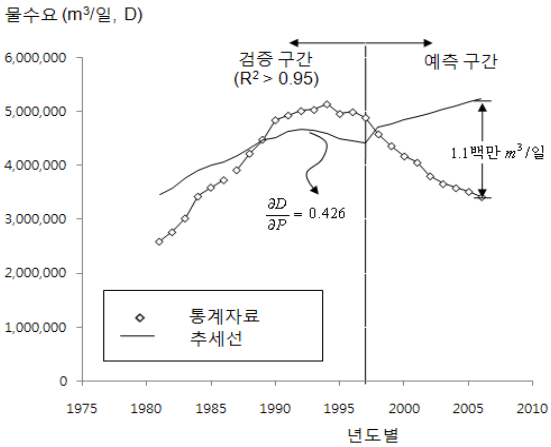


Fig. 2. 서울시 자료를 이용한 1인당원단위 모형의 예측 예

2. 연구 대상 및 범위

이러한 상황은 모형이 통계적으로 신뢰할 수 있도록 물수요를 묘사하는 것과 최근의 추세에 효과적으로 적응하는 것이 서로 모순적일 수 있음을 의미한다. 앞선 서울시의 예를 들어, 1994년 이전의 오래된 물수요 자료를 이용하는 것은 불필요하게 현재와 다른 추세를 고려하게 되므로 상수도 시설 계획상의 적응력을 떨어뜨리게 됨은 분명하다. 즉, 최근의 자료들을 근거로 추세를 신뢰할 수 있게 예측할 수 있는 모형이 필요하다. 그러나 이전의 물수요 자료를 이용하지 않을 경우에는 측정 횟수의 부족으로 통계적으로 신뢰할 수 있는 모형을 만들기까지 너무 많은 시간이 필요하다 (익명의 한 심사위원은 국내 물수요추정시 최소한 10년에서 20년 이상의 자료가 사용되고 있음을 언급한 바 있다). 또한 그 동안 도시는 다른 패턴으로 성장할 수 있으며, 수요관리와 같이 물이용과 관련된 다양한 정책의 변화로 또 다른 추세

상하수도학회지, 논문
23권 2호, pp.241-249, 4월, 2009

변화를 겪을 가능성이 높다. 이러한 상황에서 이론적으로나 실용적으로 가장 효과적인 접근법은 오래된 자료에 의존하지 않고 최근 추세에 대한 학습속도를 증가시킬 수 있는 모형을 얻는 것이다 (이상은, 2008; 이상은 등, 2009; Gleick, 2003; Lee *et al.*, 2008). 물론 물수요의 과거 자료들이 더 이상 직접적으로 도움이 되지 않는 상황에서, 초기에는 물수요에 대한 인식이 불확실하거나 주관적인 수준에 머무를 수밖에 없다. 그러나 최소한의 샘플 자료에 근거하여 시계열 모형의 매개변수들에 대한 사후적인 판단에 성공할 수 있다면 모형은 수요의 변화에 대한 학습속도를 크게 증가시킬 수 있다. 이러한 관점으로부터 모형의 가치는 신뢰할 수 있는 예측을 수행하기 위해 얼마나 적은 수의 샘플이 필요한지에 달려있다고 말할 수 있다.

이미 물수요 예측을 위한 다양한 모형이 개발되어 있기 때문에 추가적인 모형을 구현하는 연구는 소모적인 활동으로 비춰질 수 있다. 그러나 기 개발된 대부분의 모형들이 과거의 자료를 얼마나 잘 설명할 수 있는가에 주요한 초점을 맞추고 있음을 고려할 때 이 모형들을 가지고 불확실한 추세에 대한 적응력을 높이기는 한계가 있다. 특히, 상수도 시설계획시 현 상황에서 중요한 것은 모형이 갑작스러운 물수요의 추세 변화를 반영할 수 있는지를 평가하는 것이나, 이를 위한 모델링 방법은 이론적으로 제대로 정립되어 있지 않은 상황이다. 본 연구에서는 하나의 대안적 시도로서, 패턴인식에 있어서 대표적인 통계추론 과정인 베이지안 매개변수 산정법 (Bayesian Parameter Estimation)을 이용하여 시계열 모형을 구현하는 절차를 제시하고자 한다. 또한 연구범위는 목표연도 2006년을 기준으로 서울시 물수요 예측의 적절성으로 한정된다.

3. 방법론: 샘플 표본집단을 이용한 시계열 모형의 베이지안 매개변수 산정법

일반적으로, 베이즈 법칙 (Bayes' theorem)은 관찰하려는 대상에 대해 사전에 알고 있던 정보의 확률과 직접 관찰한 몇 가지 수집된 정보의 확률을 결합하여 정보에 대한 사후적인 의사결정을 하는 데에 필요한 이론적 근거를 제공한다. 또한 베이즈 법칙은 전통적인 확률이론 (배도선 등, 2003) 뿐만 아니라 현대 패턴인식에서도 가장 중요한 이론 가운데에 있다 (Duda *et al.*, 2000). 직접 관찰한 정보가 샘플을 통해 수집될 때 베이즈 법칙은 다음 식(1)과 같은 관계로 표현될 수 있다.

$$h_1(\theta|y)=\frac{h_0(\theta)f(y|\theta)}{\int_\theta h_0(\theta)f(y|\theta)d\theta}=\frac{h_0(\theta)f(y|\theta)}{g(y)} \quad (1)$$

여기서, $h_0(\theta)$와 $h_1(\theta|y)$는 각각 대상 정보의 자연상태 (state of nature) θ에 대한 주관적인 사전 분포와 표본집단의 통계량 y를 고려하여 추정된 θ에 대한 사후 분포를 의미한다. 또한, $f(y|\theta)$는 표본집단의 θ에 대해 파악된 y의 우도 (likelihood)를, $g(y)$는 모든 θ에 대해 평균된 y의 비조건적 분포 (unconditional distribution) 또는 주변 분포 (marginal distribution)를 지칭한다. 시계열 자료에 대해 매개변수를 추정하는 베이지안 매개변수 산정법(Bayesian parameter estimation)을 적용하기 위해서 우선 시계열 자료는 다음 식(2)와 같이 일반화된다.

$$\begin{aligned} x(t) &= b_1 z_1(t)+b_2 z_2(t)+\ \cdots\ +b_k z_k(t)+\epsilon(t) \\ &= Z(t)'B+\epsilon(t) \end{aligned} \quad (2)$$

여기서, $(')$는 행렬의 전치 (transpose)를 지칭한다. $x(t)$와 ϵ_t는 t년도에서 대상정보의 예측값 및 무작위오차 (random error)로 $\epsilon_t \sim N(0,\sigma_\epsilon^2)$을 가정한다. 그리고 b_i와 z_i는 시계열 모형의 i번째 매개변수와 시간함수를 의미하며, $Z(t)=[z_1(t),z_2(t),\cdots,z_k(t)]'$ 그리고 $B=[b_1, b_2, \cdots, b_k]'$로 정의된 행렬이다. 매개변수 행렬 B에 대한 주관적인 사전 확률분포 $h_0(B)$에 T개의 샘플 관측치 $\hat{X}=[x(1),\ x(2),\cdots\ ,\ x(T)]'$을 통해 얻은 매개변수 추정치 $\hat{B}=[\hat{b_1},\ \hat{b_2},\cdots\ ,\ \hat{b_k}]'$의 우도 $f(\hat{B}|B)$를 고려하여 사후 확률분포 $h_1(B|\hat{B})$로 변경하려 할 때 베이즈 법칙은 식(3)과 같이 표현될 수 있다.

$$h_1(B|\hat{B})=\frac{h_0(B)f(\hat{B}|B)}{\int_B h_0(B)f(\hat{B}|B)dB}=\frac{h_0(B)f(\hat{B}|B)}{g(\hat{B})} \quad (3)$$

위의 식(3)를 이용하여 매개변수를 추정할 때에 사전 확률분포 $h_0(B)$에 대한 형태는 이미 알려져 있는 것으로 가정된다. 많은 경우 다변량 정규밀도 (multivariate normal density)가 이용되며, 이 확률밀도는 식(3)에 의한 재생산시에도 동일한 형태를 유지한다 (Raiffa and Schlaifer, 1961; Montgomery, 1976). 즉, 사후 확률 분포 $h_1(B|\hat{B})$ 역시 다변량 정규 밀도이다. 사후 확률 분포에 대한 구체적인 관계식을 얻기 위해, 먼저 추정하고자 하는 매개변수들이 다변량 정규밀도 $B \sim N(\dot{B},\dot{V})$를 만족하는 것으로 사전 지식을 갖고 있다고 가정하자. 이 때 매개변수에 대한 사전 확률 분포는,

$$h_0(B)=(2\pi)^{-\frac{k}{2}}|\dot{V}|^{\frac{1}{2}}\exp[-\frac{1}{2}(B-\dot{B})'\dot{V}^{-1}(B-\dot{B})] \quad (4)$$

마찬가지로, $Cov[B]\equiv Cov[\hat{B}]=\hat{V}$이라면, 샘플 관측치의 우도를 식(5)와 같이 정의할 수 있다.

$$f(\hat{B}|B)\equiv(2\pi)^{-\frac{k}{2}}|\hat{V}|^{\frac{1}{2}}\exp[-\frac{1}{2}(\hat{B}-B)'\widehat{V^{-1}}(\hat{B}-B)] \quad (5)$$

위 식(5)를 적용하기 위해서는 $\hat{B}$과 $\hat{V}$가 필요하다. 이를 위해, 총 T개의 샘플, 즉, $\hat{X}$을 얻었다고 하면, 식(2)를 다음 식(6)과 같이 표현할 수 있다.

$$\hat{X}=\hat{J}\hat{B}+\hat{e} \quad (6)$$

여기서,

$$\hat{X}=\begin{bmatrix}x(1)\\x(2)\\\cdot\\\cdot\\\cdot\\x(T)\end{bmatrix},\ \hat{J}=\begin{bmatrix}Z(1)'\\Z(2)'\\\cdot\\\cdot\\\cdot\\Z(T)'\end{bmatrix}=\begin{bmatrix}z_1(1)&z_2(1)&\cdots&z_k(1)\\z_1(2)&z_2(2)&\cdots&z_k(2)\\\cdots&\cdots&\cdots&\cdots\\\cdots&\cdots&\cdots&\cdots\\\cdots&\cdots&\cdots&\cdots\\z_1(T)&z_2(T)&\cdots&z_k(T)\end{bmatrix},$$

그리고 $\hat{e}=\begin{bmatrix}\epsilon(1)\\\epsilon(2)\\\cdot\\\cdot\\\cdot\\\epsilon(T)\end{bmatrix}$ 이다. 식(6)의 오차제곱합은

$$SSE=\sum_{t=1}^{T}\{\epsilon(t)\}^2=\hat{e}'\hat{e}=(\hat{X}-\hat{J}\hat{B})'(\hat{X}-\hat{J}\hat{B})$$
$$=\hat{X}'\hat{X}-2\hat{B}'\hat{J}'\hat{X}+\hat{B}'\hat{J}'\hat{J}\hat{B} \quad (7)$$

식(7)은 미분을 통해, 오차제곱합을 최소화하기 위한 $\hat{B}$의 조건으로 다음 식(8)을 얻을 수 있다.

$$\hat{B}=(\hat{J}'\hat{J})^{-1}\hat{J}'\hat{X} \quad (8)$$

또한 각 기간별 오차항은 평균 0, 분산 σ_ϵ^2의 확률변수이므로

$$E[\hat{e}]=0,\ Cov[\hat{e}]=\begin{bmatrix}\sigma_\epsilon^2&0&0&\cdots&0\\0&\sigma_\epsilon^2&0&\cdots&0\\\cdots&\cdots&\cdots&\cdots&\cdots\\0&0&0&\cdots&\sigma_\epsilon^2\end{bmatrix}=\sigma_\epsilon^2 I \quad (9)$$

식(8)과 (9)를 적용하면, 식(10)과 같이 $\hat{V}$를 산정할 수 있다.

$$\hat{V}=E[(\hat{B}-B)(\hat{B}-B)']=\sigma_\epsilon^2(J'J)^{-1} \quad (10)$$

이 때 사후 확률분포는

$$h_1(B|\hat{B})=\frac{h_0(B)f(\hat{B}|B)}{g(\hat{B})}=\alpha\times\exp[-\frac{1}{2}\{(B-\dot{B})'\dot{V}^{-1}(B-\dot{B})+(\hat{B}-B)'\widehat{V^{-1}}(\hat{B}-B)\}] \quad (11)$$

여기서, α는 $\hat{B}$에만 종속적인 임의의 계수이다. 한편 다변량 정규 밀도는 사후적으로 $B|\hat{B}\sim N(\ddot{B},\ddot{V})$를 따르기 때문에,

$$h_1(B|\hat{B})=\alpha'\times\exp[-\frac{1}{2}(B-\ddot{B})'\ddot{V}^{-1}(B-\ddot{B})] \quad (12)$$

여기서, α'도 $\hat{B}$에만 종속적인 임의의 계수이다. 식(11)과 식(12)의 지수함수내 계수비교를 통해 사후 확률분포를 산정하기 위한 식(13)과 식(14)를 얻을 수 있다.

$$\ddot{B}=\ddot{V}(\dot{V}^{-1}\dot{B}+\widehat{V^{-1}}\hat{B}) \quad (13)$$
$$\ddot{V}^{-1}=\dot{V}^{-1}+\widehat{V^{-1}} \quad (14)$$

또한 위와 같은 시계열 모형 매개변수 추정 결과들을 가지고 예측을 하고자 할 때 예측의 기대값 및 불확실성은 각각 식(15) 및 식(16)과 같이 산정할 수 있다.

$$E[x(t)]=\sum_{i=1}^{k}b_i z_i(t)=Z(t)'\ddot{B} \quad (15)$$

$$Var[x(t)]=Var[Z(t)'\ddot{B}]+\sigma_\epsilon^2$$
$$=\sum_{i=1}^{k}\sum_{j=1}^{k}z_i(t)z_j(t)Cov[\ddot{b}_i,\ddot{b}_j]+\sigma_\epsilon^2$$
$$=Z(t)'\ddot{V}Z(t)+\sigma_\epsilon^2 \quad (16)$$

이처럼 시계열 모형에 대해 베이지안 매개변수 산정법을 적용하여 장래 예측을 하는 것은 순차적으로 (1) 시계열 모형 $x(t)$의 정의, (2) 매개변수에 대한 사전 확률분포 $h_0(B)$ 산정, (3) 샘플 표본으로부터 우도 $f(\hat{B}|B)$ 산정, (4) 사후 확률분포 $h_1(B|\widehat{B})$ 산정, 그리고 (5) 특정 년도의 예측값 $E[x(t)]$ 및 불확실성 $Var[x(t)]$ 산정으로 구성된다고 할 수 있다.

상하수도학회지, 논문
23권 2호, pp.241-249, 4월, 2009

4. 모형 구현

1995년 이후 서울시 물수요가 더 이상 과거와 같이 일정하게 증가하지 않았다. 따라서 과거 자료를 이용하는 대신에 $t \geq 1995$에서 물수요 $x(t)$를 추정하는 데에 정보가 부족하며, 단지 식(17)과 같은 1차 추세선의 형태를 따른다고 가정하자.

$$x(t) = b_1 + b_2(t-1995) + \epsilon_t = Z(t)'B + \epsilon_t \quad (17)$$

단, $B = [b_1,\ b_2]'$, $Z(t) = [1,\ t-1995]'$, 그리고 $\epsilon_t \sim N(0,\ \sigma_\epsilon^2)$이다. 식(17)의 매개변수 B에 대해서 사전적으로 이용할 수 있는 객관적인 자료는 없지만, 1995년의 물수요는 4.5백만 톤/년과 5.5백만 톤/년 사이이며, 연간 증가율 또한 0.5백만 톤/년과 −0.5백만 톤/년이라고 주관적으로 가정하는 것은 과거의 경험을 볼 때 안전하다. 따라서 각 매개변수가 주어진 범위내에서 정규분포를 갖기 위해 기대값 및 분산을 식(18)~식(21)과 같이 정량화할 수 있다.

$$E[\dot{b}_1] = \frac{5.5+4.5}{2} = 5.0 \quad (18)$$

$$E[\dot{b}_2] = \frac{0.5-0.5}{2} = 0.0 \quad (19)$$

$$Var[\dot{b}_1] = (\frac{5.5-4.5}{6})^2 = \frac{1}{36} \quad (20)$$

$$Var[\dot{b}_1] = (\frac{0.5+0.5}{6})^2 = \frac{1}{36} \quad (21)$$

공분산에 대한 정보가 부족하므로 $Cov[\dot{b}_1,\dot{b}_2] = 0$으로 잠정적으로 간주하여, 매개변수에 대한 사전 확률분포 $h_0(B)$를 식(22)와 같이 이변량 정규분포의 형태로 표현할 수 있다.

$$h_0(B) = (2\pi)^{-1}|\dot{V}|^{\frac{1}{2}}\exp[-\frac{1}{2}(B-\dot{B})'\dot{V}^{-1}(B-\dot{B})] \quad (22)$$

여기서, $\dot{B} = \begin{bmatrix}\dot{b}_1\\ \dot{b}_2\end{bmatrix} = \begin{bmatrix}5.0\\ 0.0\end{bmatrix}$,

그리고 $\dot{V} = \begin{bmatrix} Var[\dot{b}_1] & Cov[\dot{b}_1,\dot{b}_2] \\ Cov[\dot{b}_1,\dot{b}_2] & Var[\dot{b}_2]\end{bmatrix} = \begin{bmatrix}0.02778 & 0\\ 0 & 0.02778\end{bmatrix}$

이다. 또한 1994년까지 1차 추세선으로 시계열 모형을 만들었을 때 무작위 오차는 $\sigma_\epsilon^2 = 0.02060$로 산정되었으며, 이후에도 동일한 형태의 시계열 모형은 동일한 수준의 오차를 발생시키는 것으로 가정한다.

4.1 2개의 샘플 관측치 이용시

2개의 샘플 관측치만을 학습한 시계열 모형을 산정하고 모형의 특징에 대해 검토키로 하자. 이를 위해 상수도 통계자료 (환경부, 연도별)로부터, 1995년과 1996년의 서울시 물수요 자료 $x(1995)$=4,959백만 톤/년, 그리고 $x(1996)$=4,991백만 톤/년을 식(6)에 적용하면,

$$\hat{X} = \hat{J}\hat{B} + \hat{e} \quad (25)$$

여기서, $\hat{X} = \begin{bmatrix}x(1995)\\ x(1996)\end{bmatrix} = \begin{bmatrix}4.959\\ 4.991\end{bmatrix}$,

$\hat{J} = \begin{bmatrix}Z(1995)'\\ Z(1996)'\end{bmatrix} = \begin{bmatrix}1 & 0\\ 1 & 1\end{bmatrix}$

그리고 $\hat{e} = \begin{bmatrix}\epsilon(1995)\\ \epsilon(1996)\end{bmatrix}$이다. 최소오차제곱 방정식 식(8)과 식(10)을 통해 샘플 표본으로부터 우도 $f(\hat{B}|B)$를 식(26)과 같이 산정할 수 있다.

$$f(\hat{B}|B) \equiv (2\pi)^{-1}|\hat{V}|^{\frac{1}{2}}\exp[-\frac{1}{2}(\hat{B}-B)'\widehat{V^{-1}}(\hat{B}-B)] \quad (26)$$

여기서, $\hat{B} = (\hat{J}'\hat{J})^{-1}\hat{J}'\hat{X} = \begin{bmatrix}4.959\\ 0.03200\end{bmatrix}$, 그리고 $\hat{V} = \sigma_\epsilon^2(J'J)^{-1}$

$= \begin{bmatrix}0.02060 & -0.02060\\ -0.02060 & 0.04120\end{bmatrix}$이다. 따라서 사후 확률분포 $h_1(B|\widehat{B})$는 식(13)과 식(14)에 의해 식(27)과 같이 산정된다.

$$h_1(B|\hat{B}) = (2\pi)^{-1}|\ddot{V}|^{\frac{1}{2}}\exp[-\frac{1}{2}(B-\ddot{B})'\ddot{V}^{-1}(B-\ddot{B})] \quad (27)$$

여기서, $\ddot{B} = \begin{bmatrix}4.979\\ 0.006709\end{bmatrix}$, 그리고

$\ddot{V} = \begin{bmatrix}0.009504 & -0.005457\\ -0.005457 & 0.01496\end{bmatrix}$이다. $\ddot{b}_2 = 0.006709$임은 사전 정보와 샘플 정보를 결합했을 때 물수요가 매년 0.0067백만톤 증가, 즉, 물수요의 정체를 의미한다. 식(27)로부터 만들어진 시계열 모형의 기대값 및 분산은 각각 식(28) 및 식(29)와 같다.

$$E[x(t)] = Z(t)'\ddot{B} = 4.979 + 0.006709(t-1995) \quad (28)$$

$$Var[x(t)] = Z(t)'\ddot{V}Z(t) + \sigma_\epsilon^2 = 0.01496t^2 - 59.69t + 95{,}540 \quad (29)$$

또한 시계열 모형의 95 % 신뢰구간은

$$4.979+0.006709(t-1995) \pm 1.960\sqrt{0.01496t^2-59.69t+95,540} \quad (30)$$

최종적으로 2개의 샘플자료를 이용한 시계열 모형은 **Fig. 3**와 같은 예측 결과를 만들게 된다. 2개의 샘플 자료만을 이용할 경우 95 % 신뢰구간이 너무 넓어 물수요가 빠르게 증가하는 예측과 빠르게 감소하는 예측을 모두 포괄하고 있다. 즉, 예측시 불확실성이 너무 커 구현된 모형에 대해 통계적으로 그 유용성을 언급하기 힘들다.

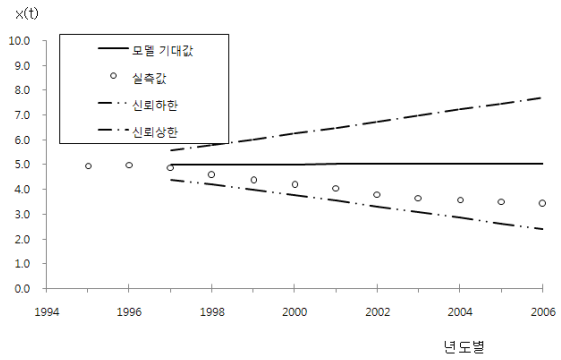


Fig. 3. 2개의 샘플 자료 이용시 시계열 모형의 물수요 예측

4.2 4개의 샘플 관측치 이용시

다음으로는 4개의 샘플 관측치를 학습한 시계열 모형을 산정하고 모형의 특징에 대해 검토키로 하자. 즉, 1997년과 1998년의 서울시 물수요 자료가 $x(1997)=4.886$ 백만 톤/년, 그리고 $x(1998)=4.580$ 백만 톤/년을 추가적으로 알게 된 경우이다. 이때 베이즈 법칙의 재귀성 (Duda *et al.*, 2000)으로부터 식(27)의 결과는 사전 확률분포 $h_0(B)$로 이용될 수 있다. 위와 동일한 방법으로 두 샘플 정보가 더 추가된다고 할 때,

$$\hat{X}=\hat{J}\hat{B}+\hat{e} \quad (31)$$

여기서, $\hat{X}=\begin{bmatrix}x(1997)\\x(1998)\end{bmatrix}=\begin{bmatrix}4.886\\4.580\end{bmatrix}$,

$\hat{J}=\begin{bmatrix}Z(1997)'\\Z(1998)'\end{bmatrix}=\begin{bmatrix}1 & 2\\1 & 3\end{bmatrix}$, 그리고 $\hat{e}=\begin{bmatrix}\epsilon(1997)\\\epsilon(1998)\end{bmatrix}$

이다. 또한 샘플 표본으로부터 우도 $f(\hat{B}|B)$는

$$f(\hat{B}|B)\equiv(2\pi)^{-1}|\hat{V}|^{\frac{1}{2}}\exp[-\frac{1}{2}(\hat{B}-B)'\hat{V}^{-1}(\hat{B}-B)] \quad (32)$$

여기서, $\hat{B}=(\hat{J}'\hat{J})^{-1}\hat{J}'\hat{X}=\begin{bmatrix}5.498\\-0.3060\end{bmatrix}$,

그리고 $\hat{V}=\sigma_\epsilon^2(J'J)^{-1}=\begin{bmatrix}0.2678 & -0.1030\\-0.1030 & 0.04120\end{bmatrix}$이다. 따라서 사후 확률분포 $h_1(B|\hat{B})$는

$$h_1(B|\hat{B})=(2\pi)^{-1}|\ddot{V}|^{\frac{1}{2}}\exp[-\frac{1}{2}(B-\ddot{B})'\ddot{V}^{-1}(B-\ddot{B})] \quad (33)$$

여기서, $\ddot{B}=\begin{bmatrix}5.0110\\-0.1060\end{bmatrix}$, 그리고 $\ddot{V}=\begin{bmatrix}0.008958 & -0.003646\\-0.003649 & 0.002881\end{bmatrix}$

이다. $\ddot{b}_2=-0.1060$임은 사전 정보와 샘플 정보를 결합하였을 때 물수요가 매년 0.106백만 톤 감소함을 의미한다. 식(33)으로부터 만들어진 시계열 모형의 기대값 및 분산은 식(34) 및 식(35)와 같다.

$$E[x(t)]=Z(t)'\ddot{B}=5.0110-0.1060(t-1995) \quad (34)$$

$$Var[x(t)]=Z(t)'\ddot{V}Z(t)+\sigma_\epsilon^2 = 0.002881t^2-11.50t+11,470 \quad (35)$$

또한 시계열 모형의 95 % 신뢰구간은

$$5.0110-0.1060(t-1995) \pm 1.960\sqrt{0.002881t^2-11.50t+11,470} \quad (36)$$

최종적으로 4개의 샘플자료를 학습한 시계열 모형의 예측결과는 **Fig. 4**와 같다. 95 % 신뢰구간내 물수요의 예측값의 범위는 다소 줄어들었고 실측값이 그 범위 안에 포함되어 있음을 알 수 있다. 그러나 모형의 학습부족으로 인해 기대값은 실측치에 대해 위로 편향되어 있으며, 물수요가 감소할 것인지, 아니면 정체할 것인지에 대해 확실한 판단을 하기는 여전히 힘들다. 따라서 4개의 샘플자료만을 학습하여 모형을 구현할 때에도 예측 불확실성이 높아 모형의 유용성을 확신하기는 힘들다고 판단된다.

4.3 6개의 샘플 관측치 이용시

1999년과 2000년의 서울시 물수요 자료가 $x(1999)=4.361$ 백만 톤/년, 그리고 $x(1998)=4.171$ 백만 톤/년을 추가하여, 6개의 샘플 관측치를 학습한 시계열 모형을 산정키로 하자. 위와 동일한 방법을 적용하여 얻은 사후 확률분포 $h_1(B|\hat{B})$는 식(37)과 같다.

상하수도학회지, 논문
23권 2호, pp.241-249, 4월, 2009

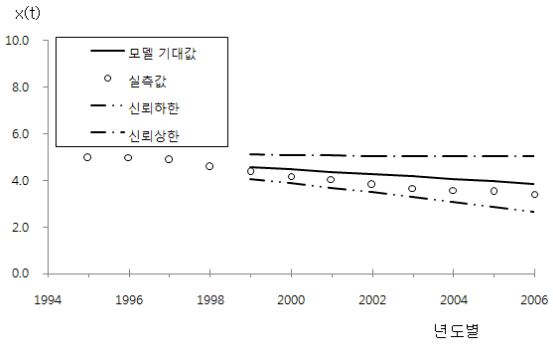


Fig. 4. 4개의 샘플 자료 이용시 시계열 모형의 물수요 예측

$$h_1(B|\hat{B}) = (2\pi)^{-1}|\ddot{V}|^{\frac{1}{2}}\exp[-\frac{1}{2}(B-\ddot{B})'\ddot{V}^{-1}(B-\ddot{B})] \quad (37)$$

여기서, $\ddot{B}=\begin{bmatrix}5.057\\-0.1624\end{bmatrix}$,

그리고 $\ddot{V}=\begin{bmatrix}0.007615 & -0.002049\\-0.002049 & 0.000921\end{bmatrix}$ 이다. $\ddot{b}_2=-0.1624$ 임은 사전 정보와 샘플 정보를 조율한 결과 물수요가 매년 0.162백만 톤 감소를 의미한다. 식(37)으로부터 만들어진 시계열 모형의 기대값 및 분산은 식(38) 및 식(39)와 같다.

$$E[x(t)] = Z(t)'\ddot{B} = 5.057 - 0.1624(t-1995) \quad (38)$$

$$Var[x(t)] = Z(t)'\ddot{V}Z(t) + \sigma_\epsilon^2 = 0.000921t^2 - 3.675t + 3,665 \quad (39)$$

또한 시계열 모형의 95 % 신뢰구간은

$$5.057 - 0.1624(t-1995) \pm 1.960\sqrt{0.000921t^2 - 3.675t + 3,665} \quad (40)$$

최종적으로 6개의 샘플자료를 학습한 시계열 모형의 예측 결과는 Fig. 5와 같다. 95 % 신뢰구간내 물수요의 예측값의 범위는 충분히 좁아서 추세 및 물수요 감소 크기에 대한 판단이 가능하다. 또한 예측 기대값이 실측 자료에 대해 편향되지 않고 잘 설명하고 있음을 알 수 있다. 따라서, 베이지안 매개변수 산정법을 이용하여 6개의 샘플자료에 대한 학습 결과 새로운 추세에 적응하고 통계적으로 유용한 시계열 모형을 만들 수 있는 것으로 판단된다.

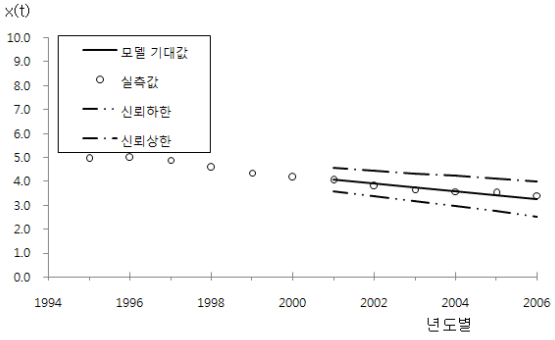
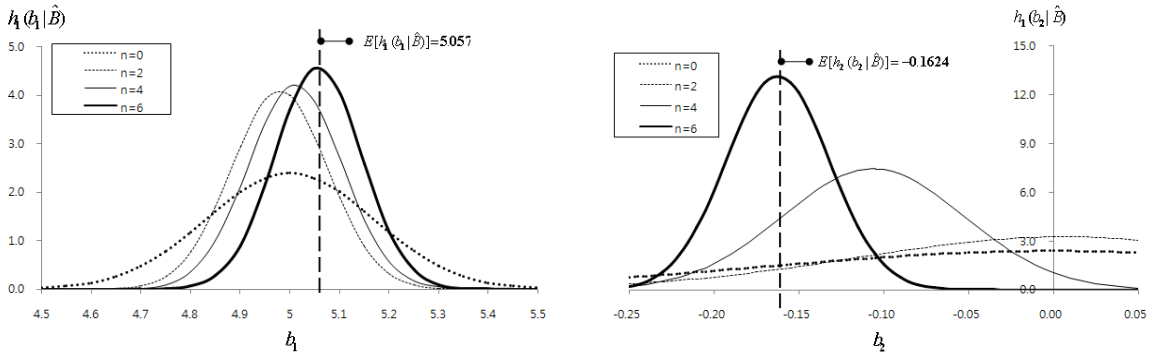


Fig. 5. 6개의 샘플 자료 이용시 시계열 모형의 물수요 예측



(a) b_1에 대한 매개 변수 추정

(b) b_2에 대한 매개 변수 추정

Fig. 6. 매개변수 추정 결과

5. 결과 및 검토

서울시 물수요에 대한 시계열 모형으로서 두 매개변수 b_1과 b_2로 구성된 1차 추세선을 구현하였다. 초기 물수요 추세의 변화에 의해 각 매개변수에 대한 사전 정보는 매우 미약하였다. 따라서 **Fig. 6**와 같이, 주관적으로만 정량화할 때, 즉 $n=0$에서 각 매개변수의 확률분포는 매우 큰 분산 또는 불확실성을 갖고 있다. 그러나 샘플 정보가 추가될수록 각 확률분포는 기대값을 중심으로 분산이 빠르게 줄어든다. 특히, 연간 물수요 증가량에 해당하는 b_2에 대한 확률은 샘플 수가 증가할수록 큰 변화를 보였다. 이는 샘플수에 따라 물수요 추세에 대한 학습효과가 큼을 의미한다.

또한 샘플 표본수에 따라 2006년 물수요량 예측 결과는 **Fig. 7**과 같이 변화된다. 샘플이 전혀 없을 때와 샘플이 2개만을 이용할 경우 기대값이 약 5백만 톤이었지만 불확실성이 매우 높기 때문에 예측 결과에 통계적 의미를 두는 것은 부적절하다. 그러나 샘플수가 증가하여 학습정도가 늘어남에 따라서 불확실성은 크게 감소함을 알 수 있다. 특히, 6개의 샘플 표본을 이용한 경우 2006년의 물수요는 기대값 3.27백만 톤/년을 중심으로 뚜렷한 피크를 형성하고 있으며 실측치 3.41백만 톤/년과 매우 유사한 값을 보이고 있다.

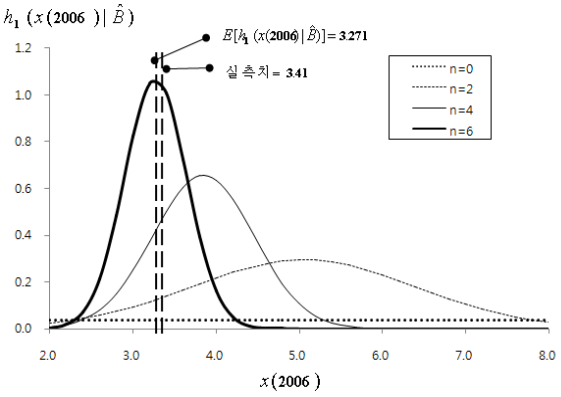


Fig. 7. 2006년 물수요량 예측 결과

위의 결과들은 서울시 물수요 예측시 동 방법론을 적용한다면 최근 6년의 샘플 표본을 가지고 신뢰할 수 있는 시계열 모형을 만들 수 있음을 의미한다. 즉, 물수요 추세가 변한 직후 미래 물수요에 대해 신뢰할 수 있는 예측이 불가능하지만, 6년 이후 새로운 추세에 대한 학습으로 다시금 신뢰할 수 있는 시계열 모형을 만들 수 있게 된다.

6. 결 론

최근 들어 대규모 상수도 시스템들의 물수요 패턴이 변화해 불필요한 용량 증설에 대한 논란이 계속되고 있다. 이러한 상황은 상수도 시설 계획시 더 많은 자료를 이용하는 것이 수요 예측에 더 도움이 되지 않음을 보이며, 나아가 상수도 시설에 대한 적응형 관리방안이 필요함을 시사한다. 과거 효율성 달성에 실패한 가장 기술적인 이유는 새로운 물수요 추세에 대한 적응 실패라고 할 수 있다. 이에 본 연구는 오래된 자료에 의존하지 않고 최근 추세에 대한 학습속도를 증가시킬 수 있는 모형을 구현할 수 있는 모델링 방법을 제시하고자 하였다. 모형이 새로운 추세에 대한 빠른 학습속도를 갖도록 대표적인 패턴인식 방법에 해당하는 베이지안 매개변수 산정법을 서울시의 물수요 예측에 적용하였다. 그 결과, 물수요가 새롭게 변화하더라도 6년이 경과된 이후에는 다시 신뢰할 수 있는 수요예측이 가능함을 확인하였다.

본 연구에 제시된 모형 및 모델링 절차는 새로운 추세에 대한 학습능력을 높이려는 필요로 인한 초기 연구로서 실제 상수도시설 계획에 적용되기 위해 다음과 같은 추후 연구가 필요할 것으로 생각된다. 첫 번째로, 본 연구에서는 물수요의 시계열 모형이 1차 추세선을 따른다고 가정하였다. 하지만 도시가 충분히 성장한 이후부터 물소비 패턴은 도시 인구 성장 외에도 생활 방식, 기후, 거시경제 상황, 도시 물관리정책 등 다양한 요인에 의해 영향을 받게 되는 것으로 잘 알려져 있다. 따라서 모형이 불확실성을 극복하고 정책적인 의미를 얻는 데에 크게 유용하기 위해서는 베이지안 매개변수 산정을 위한 시계열 모형을 1차 추세선이 아닌 다중 회귀식의 형태로 가정할 필요가 있을 것으로 판단된다. 두 번째로, 본 연구에서는 추세 변화에 확신을 할 수 없는 기간, 즉 6년 이내에는 어떠한 의사결정을 해야 하는지에 대한 언급을 하지 않았다. 본 연구자들의 견해로는, 이 의사결정을 이론적으로 만족시키기 위해서 물수요의 추세에 대한 확률 외에도, 판단으로 인한 행동 (예를 들어 정수장 용량 증설 유무)이 야기하는 손실 리스크들을 고려하여, 베이즈 판정이론을 적용시킬 필요가 있을 것으로 생각된다. 셋째, 본 연구는 서울시라는 한 사례만을 다루었다. 따라서 '베이지안 매개변수 산정을 이용할 때 새로운 패턴에 대한 6년의 자료만 확보된다면 통계적으로 적절하게 물수요 추세 변화를 인식할 수 있다'라는 연구 결과를 국내 다른 지역에 일반화하려는 것은 적절하지 않다. 따라서 구현된 방법과 모형이 상수도 계획시 보다 유용하기 위해서 사례 연구를 늘이고 지역별 특성에 대한 분석이 필요하다고 판단된다.

상하수도학회지, 논문
23권 2호, pp.241-249, 4월, 2009

사 사

본 연구는 국토해양부가 주관하고 한국건설교통기술평가원이 시행하는 2007년도 첨단도시개발사업(과제번호:07도시재생B04) 지원 사업으로 이루어진 것으로 이에 감사를 드립니다.

참고문헌

1. 감사원 (2005) 감사원 보도자료 – 상수도 개발 및 운영실태 감사 결과, 12월 15일자.
2. 국회예산정책처 (2008) *상수도 개발 및 운영 실태 평가*.
3. 박희경, 이상은, 김성훈, 신은허, 최동진 (2007) *수도시설 적정 용량 산정을 위한 적용기준 연구*, KAIST, 건설교통부•한국수자원공사.
4. 배도선 등 (2003) *통계학 이론과 응용*, 청문각, 서울.
5. 이상은 (2008) 물 인프라의 조직화된 복잡성을 다루기 위한 적응형 설계, 박사학위논문, *KAIST 건설 및 환경공학과*.
6. 이상은, 박희경 (2009) 국내 정수장 과다시설용량 실태 분석, *상하수도학회지*, 23(1), pp.57-67.
7. 환경부 (연도별) *상수도 통계*.
8. Duda, R.O., Hart, P.E., Stork, D.G. (2001) *Pattern Classification,* 2nd Edition, Wiley-Interscience, Inc., N.Y..
9. Gleick, P.H. (2003) Global freshwater resources: soft-path solutions for the 21st century, *Science*, 302(5650), pp.1524-1527.
10. Lauria, D.T. (1983) Research needs for capacity planning, *Journal of the American Water Works Association,* 75(1), pp.14-19.
11. Lee, S., Shin, J., and Park, H. (2008) Adaptive design for water infrastructures, *Proceeding of International Seminar on Civil & Infrastructure Engineering 2008*, Shah Alam, Malaysia, D6.
12. Montgomery, D.C., Johnson, L.A. (1976) *Forecasting and Time Series Analysis*, McGraw-Hill, N.Y..
13. Raiffa, H., and Schlaifer (1961) *Applied Statistical Decision Theory*, Harvard University Press, Mass..
14. Ritzman, L.P. and Krajewski, L.J. (2003) *Foundations of Operations Management*, Prentice Hall.