大 韓 土 木 學 會 論 文 集
第29卷 第4B號 · 2009年 7月
pp. 319 ~ 328

확률강우량 산정을 위한 EDA 기법의 적용

Application of EDA Techniques for Estimating Rainfall Quantiles

박현근* · 오세정** · 유철상***
Park, Hyunkeun · Oh, Sejeong · Yoo, Chulsang

Abstract

This study quantified the data by applying the EDA techniques considering the data structure, and the results were then used for the frequency analysis. Although traditional methods based on the method of moments provide very sensitive statistics to the extreme values, the EDA techniques have an advantage of providing very stable statistics with their small variation. For the application of the EDA techniques to the frequency analysis, it is necessary to normalization transform and inverse-transform to conserve the skewness of the raw data. That is, it is necessary to transform the raw data to make the data follow the normal distribution, to estimate the statistics by applying the EDA techniques, and then finally to inverse-transform the statistics of transformed data. These statistics decided are then applied for the frequency analysis with a given probability density function. This study analyzed the annual maxima one hour rainfall data at Seoul and Pohang stations. As a result, it was found that more stable rainfall quantiles, which were also less sensitive to extreme values, could be estimated by applying the EDA techniques. This methodology may be effectively used for the frequency analysis of rainfall at stations with especially high annual variations of rainfall due to climate change, etc.

Keywords : *frequency analysis, EDA, rainfall quantile, data transform*

요 지

본 연구에서는 자료의 구조를 이용하는 통계방법인 EDA 기법을 적용하여 자료를 정량화 하고, 이를 이용하여 빈도해석을 실시하였다. 모멘트법을 이용하는 전통적 방법이 극치값에 민감하게 반응하는 통계치를 주지만, EDA 기법은 변동이 적은 안정적인 통계치를 주는 장점이 있다. 빈도해석에 EDA 기법를 적용하는 경우에는 자료의 왜곡도를 반영하기 위해 원자료의 정규화 변환 및 역변환 과정을 거쳐야 한다. 즉, 원자료를 정규화 변환하고, EDA 기법을 적용하여 변환된 자료의 통계치를 추정하며, 이를 다시 역변환하여 원자료의 통계치를 결정해야 한다. 이렇게 결정된 통계치는 주어진 확률밀도함수를 이용한 빈도해석에 적용된다. 본 연구에서는 서울 및 포항지점의 연최대치 1시간 강우자료를 대상으로 분석을 수행하였다. 그 결과 EDA 기법을 적용하는 경우 극치값에 덜 민감한 안정적인 확률강우량의 산정이 가능한 것으로 확인되었다. 이러한 방법론은 특히 기후변화 등의 원인으로 강수자체의 경년변동이 매우 큰 지점의 빈도해석에 유용하게 사용될 수 있을 것이다.

핵심용어 : 빈도해석, EDA, 확률강우량, 자료변환

1. 서 론

최근 들어 세계 각국에서는 이상기상으로 인한 자연재해의 발생이 잦아지고 있다. 1993년 미국 미시시피 강의 대홍수로 인해 150억불의 피해가 기록되었고, 1998년 중국 양쯔강 유역의 대홍수로 막대한 재산과 인명피해를 발생시켰다. 유럽지역에서는 라인 강 상류의 홍수피해와 네덜란드의 홍수피해가 있었으며, 아시아 지역에서는 네팔, 인도의 홍수로 막대한 피해를, 그리고 남미 베네수엘라의 대홍수도 기록적인 인명과 재산피해를 남겼다(안재현 등, 2000; Coles *et al*., 2003). 이러한 이상기상의 원인은 대체로 지구온난화에 의한 기후변화일 것으로 받아드려지고 있다.

우리나라에서도 가뭄, 홍수 등의 기상이변이 증가하고 있는 것으로 보고되고 있다(한화진 등, 2005). 특히 1990년대 후반 들어 국지적인 소나기성 강우가 많이 발생하고 있으며, 그 빈도는 급격히 증가하는 것으로 나타났다(차은정과 최영진, 2000). 또한 1994~1995년의 가뭄, '95, '96, '98, '99년의 홍수 등과 같은 극치 기상현상이 잦아지고 있다(한국건설기술연구원, 2000). 지난 2002년 8월 31일에 태풍 루사는 하루 동안 강릉지역에 871 mm에 달하는 비를 뿌리면서 최악의 집중호우로 기록되었다(국립방재연구소, 2002).

이러한 이상기상의 빈번한 발생은 발생자체에 따른 피해뿐

*정회원 · 유량조사사업단 연구원 · 공학석사 (E-mail : gusroot@kict.re.kr)
**(주)하존이앤씨 과장 · 공학석사 (E-mail : waterkorea@korea.ac.kr)
***정회원 · 교신저자 · 고려대학교 건축 · 사회환경공학과 교수 · 공학박사 (E-mail : envchul@korea.ac.kr)

만 아니라 피해의 복구, 홍수예경보 시스템을 포함한 재해의 대비, 하천관리를 포함한 국토의 관리 등에 영향을 미치게 된다. 특히, 과거 강우과정의 정상성에 근거한 빈도해석 결과를 기준으로 한 기준설정에 큰 혼선을 가져올 가능성이 크다. 이는 2003년 태풍 루사의 경험에서 이미 확인된 바 있다. 즉, 태풍 루사시의 일최고강우량은 가능최대홍수량(probable maximum precipitation: PMP)에 근접하는 호우기록으로 이 사상을 포함하느냐 배제하느냐에 따라 24시간 100년 빈도 확률강우량에 무려 200 mm이상의 차이가 발생하는 것으로 나타난다(박상덕, 2002). 따라서 관련 주요 피해지역인 강릉, 양양, 대관령 등 강원도 동쪽 지역의 유역종합치수계획 및 하천정비기본계획의 기준설정에 큰 혼란이 야기될 수밖에 없는 상황이다.

이러한 문제점은 근본적으로 현재의 빈도해석 방법이 이상치 정도의 큰 값에 과민하게 반응하기 때문이다. Ahn *et al.*(2003)의 연구에서도 살펴본 것처럼, 이상치의 발생은 확률강우량의 큰 변화를 야기하게 된다. 그러나 이후 정상적인 규모 또는 그 이하의 년 최대치가 수년이상 기록되게 되면 그 영향은 점차 사라져, 궁극적으로 정상적이라고 판단되는 규모의 확률강우량으로 수렴하는 현상을 보인다. 과거 관측자료에 나타난 확률강우량의 변동폭은 대체로 기후학적 강우의 변동폭(climatological rainfall variability band) 이내인 것으로 파악되고 있다(Ahn *et al.*, 2003). 그러나 이러한 결론은 강우의 과정이 정상적이라고 판단할 수 있는 경우에 해당되는 결론이다. 만일 강우의 정상성 과정이 깨진다면 확률강우량은 상승 또는 하강의 추세를 보이게 된다. 물론 확률강우량 자체의 정상성 판단도 방법론의 제한(유철상 등, 2007) 및 관측기록의 한계 등(정성인 등, 2007)으로 판단이 쉽지 않은 측면이 있다.

그러나 결과적으로 보면 강우의 정상성이 가정되든 또는 확률강우량의 비정상적인 변화가 나타나던 간에 관측자료에 근거한 확률강우량의 급격한 변화는 바람직스럽지 않다. 이미 전절에서 살펴본 바와 같이 어느 한 사상의 고려여부에 따라 수백 mm 이상의 확률강우량 차이는 큰 문제가 아닐 수 없다. 따라서 확률강우량의 정상성 가정을 의심할 만한 상황이 아닌 경우에는 정상적이라고 받아들여질 수 있는 확률강우량이 추정될 수 있어야 하고, 만일 확률강우량의 정상성이 깨지는 경우라면 그러한 장기적인 추세가 고려될 수 있는 확률강우량 추정방법이 필요하다. 이는 지구온난화로 인한 기후변화의 영향이 크게 나타나는 수문학 분야의 실무에서 중요한 문제이다.

본 연구에서는 이상치의 포함여부에 따라 큰 변동폭을 보일 수밖에 없는 전통적 확률강우량 산정방법의 문제점을 보완하기 위한 방안으로 탐색적 자료분석(exploratory data analysis: EDA) 기법의 적용성을 검토하고자 한다. EDA 기법에서 사용하는 자료의 특성화는 자료의 구조에 대한 것으로 전통적 방법이 모멘트법에 근거하여 자료자체의 크기를 고려하는 것과 대비된다. 자료 자체가 아닌 자료의 구조를 고려함으로서 이상치에 대한 영향이 크게 완화될 가능성이 크다. 본 연구에서는 먼저 EDA에 대한 특성을 자세히 살펴보고, 아울러 이를 확률강우량의 추정에 적용하는데 필요한 방법론을 제시해 보고자 한다.

2. 빈도해석

2.1 전통적 방법

수문자료의 분석절차를 확률론적 수문분석기법이라 통칭하며 특히 강우나 홍수 혹은 갈수의 발생빈도를 확률론적으로 예측하는 방법을 빈도해석이라 한다. 빈도해석법은 어떤 수문사상이 발생하는 원인과 과정 등에 관해서는 전혀 상관하지 않고 오직 어떤 크기를 가진 사상이 발생할 확률(혹은 빈도)을 결정하는 것이다. 전통적 방법에 이용되는 통계학적 수문빈도해석에서는 자료계열의 독립성과 정상성이라는 두 가지 기본가정과 자료가 동일한 분포로부터 획득되었다는 가정이 필요하다(윤용남, 1998; 정종호와 윤용남, 2003).

확률분포형의 적용을 위해서는 수문자료를 이용하여 각 확률분포에서 사용하고 있는 매개변수의 추정이 필수적이다. 매개변수 추정방법에는 모멘트법(method of monents: MOM), 최우도법(method of maximum likelihood: MLM), 확률가중모멘트법(method of probability weighted: PWM), L-모멘트법 등이 있다.

수문자료의 빈도해석을 위한 전 단계로서 확률분포형의 결정이 필요하다. 이를 위해서는 적합한 확률분포형을 결정하기 위해서 분포형 별 매개변수의 적합도 검정을 실시한 후, 최적 확률분포형을 선정하여야 한다. 적합도 검정 방법으로 x^2는 검정, K-S(Kolmogorov-Smirnov) 검정, PPCC (Probability Plot Correlation Coefficient) 검정 등이 있다. 이와 관련한 자세한 내용은 정정호와 윤용남(2003), Wang(1997) 등에서 살펴볼 수 있다.

2.2 EDA 기법

탐색적 자료분석(exploratory data analysis: EDA)은 데이터의 특징과 내제하는 구조적 관계를 알아내기 위한 기법들을 말한다(Tukey, 1977). EDA 기법은 탐색과정을 통해 얻어낸 정보를 이용하여 원자료의 특성에 대한 유용한 실마리를 찾을 수 있게 하는 기법이다. EDA에서는 자료를 어떤 툴에 적용하기 보다는 자료를 있는 그대로 보려는데 더 중점을 두고 있다. 즉, EDA에서는 자료의 구조와 특징 파악을 주 목적으로 하며, 따라서 신뢰성 있는 자료의 요약과 그래프적 기법이 많이 사용된다(허명회와 이태림, 1993).

EDA의 주요 특징으로는 저항성(resistance), 잔차(residuals), 재표현(re-expression), 현시성(revelation) 등 네 가지를 들 수 있다(백운붕과 허명회, 1987; 허명회와 이태림, 1993). 먼저, 저항성이란 자료의 부분적 변동에 너무 민감하게 반응하지 않는 것을 의미한다. 자료의 일부가 기존의 자료와 현격히 다른 값으로 대체되어도 크게 다른 결과를 만들어 내지 않는다. EDA에서는 자료의 대부분을 구성하는 자료들에 관심을 기울이며, 형태가 특이한 일부의 자료들에는 관심을 두지 않는다. EDA에서는 표본평균(sample mean) 보다는 표본중위수(sample median)가 더 선호된다.

잔차는 모형 적합값을 빼고 남은 나머지를 말한다. 자료가 $(x_i,\ y_i)$로 이루어진 경우 직선 $y_i = a + bx_i$로 적합화 되었다면 잔차는 $r_i = y_i - \bar{y}_i$이다. EDA에서는 잔차의 검토작업 없이는 어떠한 자료 분석도 완료된 것으로 간주하지 않는다. 보통과 다른 잔차가 포착되었을 경우에는 해당하는 관측값

	줄기	잎		줄기	잎
54	5×10	+4		2	3
67	6×10	+7	→	3	
55	5×10	+5		4	
23	2×10	+3		5	45
				6	7

그림 1. 줄기그림의 작성예

이 어떻게 얻어졌고 다루어졌는지 면밀히 검토하여야 하고, 곡선성(curviliniarity), 비가법성(non-additivity), 분산의 이질성(heterogeneity) 등을 고려해야 한다.

재표현은 자료 분석과 해석을 단순화 할 수 있도록 적당한 척도로 바꾸는 일련의 과정이다. 즉, 특정 척도가 선호되는 경우를 제외하고 원래 측정 척도를 다른 척도로 재표현할 때 대칭성(symmetricity), 분산의 동질성(homogeneity), 직선화(linearity), 가법성(additivity) 등 자료 구조파악과 해석에 도움을 얻을 경우가 많다. 변환이라는 단어와 혼용되기도 한다.

마지막으로 현시성은 그래프로 표현하는 방법과 같이 자료에 관련된 정보의 효과적인 전달을 뜻한다. 대표적인 예로 줄기그림을 들 수 있다.

2.3 분위수 및 EDA에서의 주요 통계치 추정

다섯 숫자 요약 기법은 EDA에서 자료를 요약, 기술하는데 사용하는 방법이다. 자료를 크기순으로 나열했을 때 가운데 값은 중위수(median)가 된다. 그리고 가운데 값을 중심으로 양방향의 수가 나열되는데 이 수를 또다시 반씩 나누게 되어 얻는 숫자를 경첩(hinge)이라 한다. 또한 나열된 수의 가장 큰 수를 최대값(maximum)이라 하고 가장 작은 수를 최소값(minimum)이라 한다.

위의 숫자들에서 굵게 표시된 숫자들이 이 자료를 요약, 대표하는 값이 된다. 중앙값(median)은 자료를 크기순으로 늘어놓았을 때 가장 중앙에 위치하게 되는 자료를 뜻한다. EDA분석에서는 중앙값이 자료 분석에 있어서 중요한 인자라 할 수 있다. 만약 자료의 수를 N이라고 가정을 한다면 중앙값의 깊이(depth)는 아래와 같이 결정된다.

N이 홀수일 경우 $(N+1)/2$번째 자료점
N이 짝수일 경우 $N/2$번째와 $(N/2+1)$번째 자료점의 평균

탐색적 자료분석에서 통계치를 구할 때는 두 사분위수(upper quartile, lower quartile)를 많이 이용한다. 그것은 사분위수가 더 강한 저항성(resistance)을 가지기 때문이다. 자료의 수를 N이라고 하면 경첩(hinge, quartile)의 깊이 $d(H)$는 $d(H)=[(N+1)/2+1]/2$로 주어진다. 일반적으로 평균, 분산과 같은 통계치는 저항성(resistance)이 없으므로 EDA에서는 평균(μ_1)과 표준편차(σ)를 저항성이 강한 통계치로 대체시킨다. EDA에서의 평균과 표준편차는 아래의 식과 같이 정의된다.

$$\mu_1 = \frac{1}{2}(H_L + H_U) \tag{1}$$

$$\sigma = \frac{1}{1.349}(H_U - H_L) \tag{2}$$

여기서 H_L은 자료의 아래경첩(lower hinge) 즉 아래 사분위수(lower quartile)를 말하며, H_U는 위 사분위수(upper quartile), M은 중앙값을 말한다. 분산의 추정치에서 두 사분위수의 차이를 1.349로 나눈 이유는 평균이 μ, 분산이 σ^2인 정규분포에서 아래 사분위수와 위 사분위수가 각각 $\mu-0.6745\sigma$와 $\mu+0.6745\sigma$로서 그 차이가 1.349σ이기 때문이다. 또한 사분위수, 중위수를 이용하여 왜곡도(μ_3)를 다음 식을 이용하여 구할 수 있다.

$$\mu_3 = \frac{(H_U - M) - (M - H_L)}{(H_U - M) + (M - H_L)} \tag{3}$$

이렇게 정의된 왜곡도는 -1과 +1사이의 값을 가지며, +1에 가까울수록 오른쪽으로 꼬리를 길게 뻗는 분포임을 말하며, -1에 가까울수록 왼쪽으로 꼬리를 길게 뻗는 분포임을 말한다.

EDA 기법을 적용한 빈도해석은 모멘트법에 근거한 전통적 방법과 비교하여 기본 통계치 산정에 있어 그 차이를 보이게 된다. 즉, EDA 기법 적용 시 자료의 평균은 위 사분위수와 아래 사분위수의 산술평균으로 나타내고, 분산은 두 사분위수의 차이를 1.349로 나눈 값으로 계산한다. 또한, 사분위수, 중위수를 이용하여 Eq. (3)과 같이 왜곡도(skewness)를 계산하게 된다. 일단 이와 같이 통계치가 결정되면 이하의 과정은 전통적인 빈도해석 방법과 동일하다. 그러나 EDA 기법을 적용하여 추정된 자료의 통계치가 모멘트법을 적용하여 추정한 통계치와 유사한지에 대해서는 추가의 검토가 필요하다.

3. 자료의 수집 및 기본 통계특성

본 연구에서 서울 및 포항 2개 지점의 강우관측소를 대상으로 하였다. 먼저, 서울지점은 1961~2005년까지 45개년의 자료를 사용하였다. 그림 3(a)는 서울지점의 1시간 최대강우량의 변화를 나타낸 것이다. 45개년 1시간 최대강우의 평균은 44.5 mm이다. 특히, 1964년에는 116.0 mm의 극치 강우사상이 관측되었고, 1966년(73.0 mm), 2001년(90.0 mm)에도 평균값을 크게 상회하는 강우가 관측되었다. 또한 그림 3(b)는 서울지점의 24시간 최대강우량의 변화를 나타낸 것이

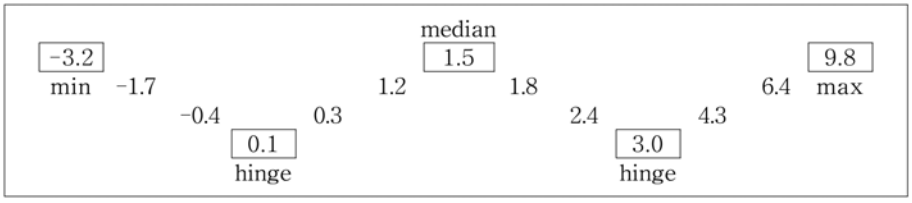
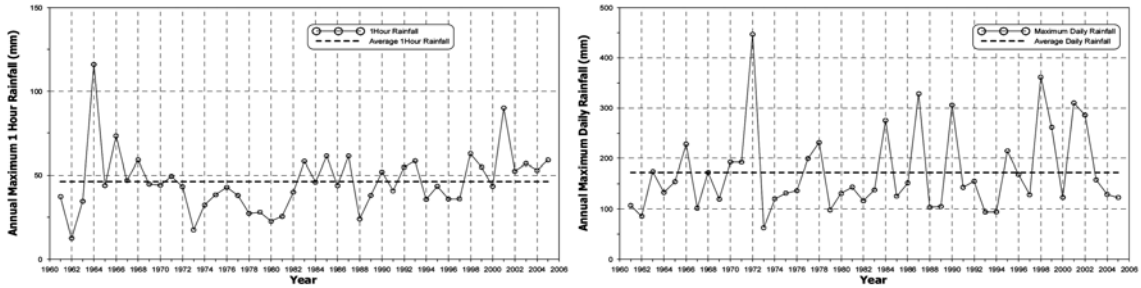
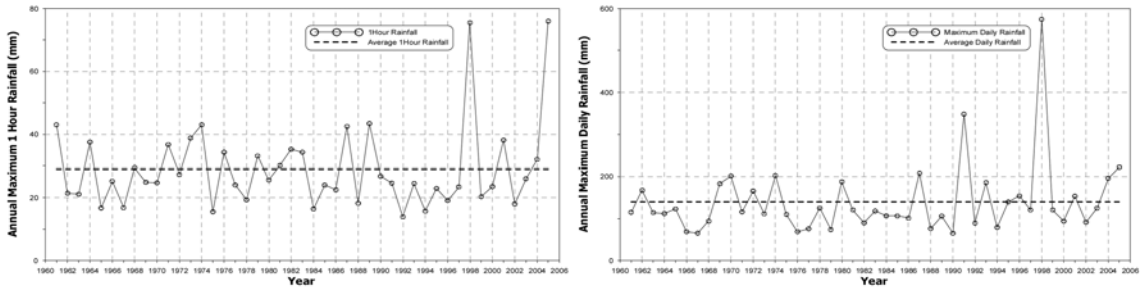


그림 2. 다섯숫자 요약의 예



(a) 1시간 (b) 24시간

그림 3. 지속시간 별 서울지점 최대강우량의 변화



(a) 1시간 (b) 24시간

그림 4. 지속시간 별 포항지점 최대강우량의 변화

다. 45개년 24시간 최대강우의 평균은 172.5 mm이다. 특히, 1972년에 446.8 mm의 극치 강우사상이 관측되었고, 1987년(328.4 mm), 1990년(305.8 mm), 1998년(361.5 mm)에도 평균값을 크게 상회하는 강우가 관측되었다.

포항지점의 경우도 사용된 자료기간은 1961~2005년까지로 동일하다. 그림 4(a), (b)는 포항지점의 1시간 및 24시간 최대강우량의 변화를 나타낸 것이다. 1시간 최대강우의 평균은 29.0 mm이고, 24시간 최대강우의 평균은 139.4 mm이다. 1시간 자료의 경우 1998년, 2005년에는 각각 75.5 mm와 76.0 mm의 극치 강우사상이 관측되었고, 1974년(43.0 mm), 1987년(42.5 mm), 1989년(43.4 mm)에도 평균값을 크게 상회하는 강우가 관측되었다. 24시간 자료의 경우 1998년에 574.3 mm의 극치 강우사상이 관측되었고, 1991년(347.9 mm), 2005년(220.0 mm)에도 평균값을 크게 상회하는 강우가 관측되었다.

표 1. 서울지점 연최대치 강우자료에 대한 기본 통계 특성

지속시간 (전통적 방법)	평균	분산	표준편차	왜곡도
1시간	46.3	327.7	18.1	1.4
6시간	116.7	1904.4	43.6	1.2
12시간	144.4	2806.2	53.0	1.0
24시간	172.5	6799.3	82.5	1.4

표 2. 포항지점 연최대치 강우자료에 대한 기본 통계 특성

지속시간 (전통적 방법)	평균	분산	표준편차	왜곡도
1시간	29.0	171.9	13.1	2.1
6시간	78.0	2068.4	45.4	3.6
12시간	106.3	3767.4	61.3	3.1
24시간	139.3	7344.3	85.6	3.4

표 1과 2는 서울지점 연최대치 자료에 대한 기본 통계 특성을 지속시간별로 정리한 것이다. 적정 확률분포형을 찾기 위한 적합도 검정을 수행한 결과, 서울지점의 경우 1시간 지속기간의 경우는 2변수 Gamma 분포가, 24시간 지속기간인 경우에는 2변수 Gamma 분포 및 3변수 Gamma 분포가 검정, 검정, PPCC 검정을 모두 통과하는 것으로 확인되었다. 포항지점의 경우는 1시간 지속기간의 경우는 2변수 대수정규분포, 3변수 대수정규분포, 2변수 Gamma 분포 그리고 GEV 분포가, 24시간 지속기간인 경우에는 2변수 Gamma 분포와 GEV 분포가 검정, K-S 검정, PPCC 검정을 통과하는 것으로 확인되었다.

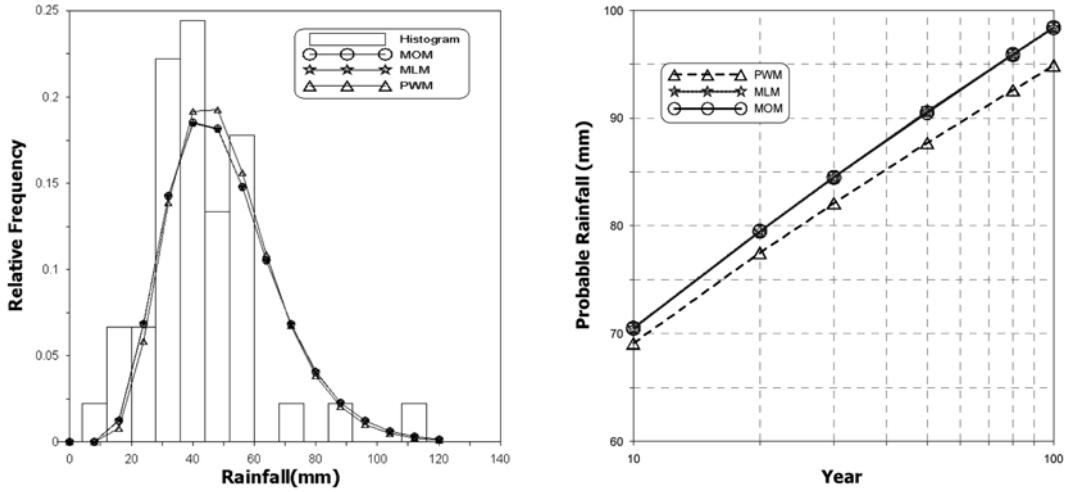
4. EDA 기법의 적용성 평가

4.1 전통적 방법을 적용한 빈도해석

그림 5(a), (b)는 서울지점(1시간 지속시간)에서 적정 확률분포형으로 결정된 2변수 Gamma 분포의 매개변수 추정방법에 따라 산정된 확률밀도함수와 확률강우량을 나타낸 것이다. 표 3은 재현기간별 매개변수 추정방법 별로 산정된 1시간 확률강우량을 비교한 것이다. 매개변수 추정방법에 따른 차이는 아주 크지는 않으며, 대략 5% 미만인 것으로 파악되었다.

4.2 관측자료에 대한 EDA 기법의 직접 적용

표 4는 서울지점의 연최대치 1시간 및 24시간 강우자료에 대해 모멘트법으로 산정된 통계치와 EDA 기법을 적용하여 산정된 통계치를 비교한 것이다. 전체적으로 EDA 기법을 적용한 경우가 모멘트법을 적용한 경우에 비해 평균은 약간 작게, 표준편차와 왜곡도는 아주 작게 추정되어 있음을 확인할 수 있었다. 이는 기본적으로 원 자료가 대칭을 이루지 못하고 있음을 의미한다.



(a) 확률밀도함수 (b) 확률강우량

그림 5. 매개변수 추정방법에 따른 2변수 Gamma 분포의 확률밀도함수와 확률강우량의 차이

표 3. 매개변수 추정방법에 따른 확률강우량의 차이(지속시간 1시간, 2변수 Gamma 분포) (단위:mm)

재현기간	MOM	MLM	PWM
10	70.5	70.5	69.1
20	79.5	79.5	77.5
30	84.5	84.5	82.1
50	90.5	90.6	87.7
80	95.9	95.9	92.6
100	98.4	98.4	94.9

표 4에 나타난 통계치를 가지고 빈도해석한 결과는 그림 6 및 표 5와 같다. 즉, 모멘트법 및 EDA 기법을 적용하여 추정한 통계치를 이용하여 2변수 Gamma 분포의 확률밀도함수를 결정하고 이를 바탕으로 확률강우량을 추정하여 비교한 것이다. 전체적으로 EDA 기법을 적용하여 산정된 확률강우량이 전통적 방법보다 확률강우량이 작게 산정되었음을 확인할 수 있었다. 이는 특히 EDA 기법을 적용한 경우의 분산 및 왜곡도가 모멘트법을 적용한 경우보다 작게 산정되기 때문이다. 이러한 원인은 EDA가 근본적으로 유사 정규분포를 가정하기 때문이기도 하다.

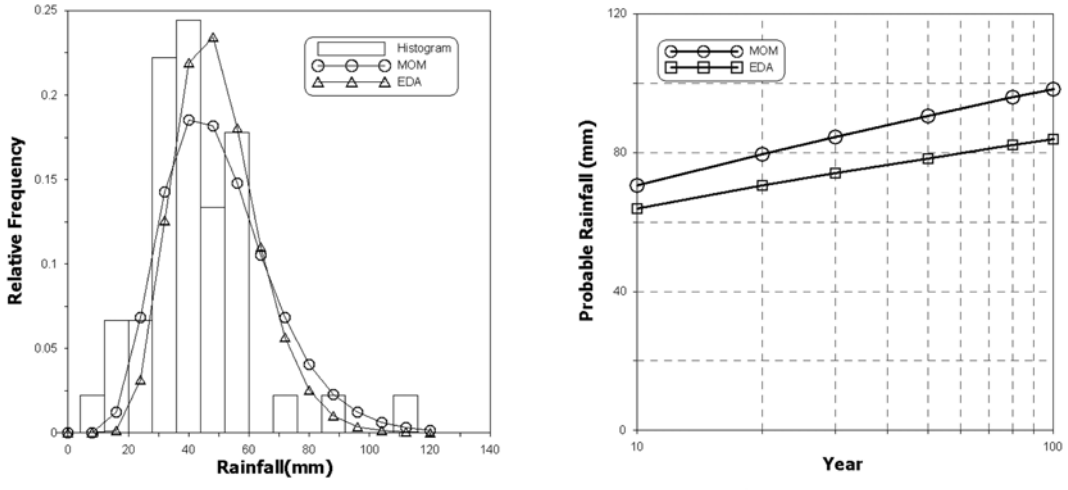
표 4. 전통적 방법과 EDA 기법을 적용하여 추정한 기본 통계치 비교

전통적 방법	평균	분산	표준편차	왜곡도
1시간	46.3	327.7	18.1	1.4
6시간	116.7	1904.4	43.6	1.2
12시간	144.4	2806.2	53.0	1.0
24시간	172.5	6799.3	82.5	1.4
EDA 기법	평균	분산	표준편차	왜곡도
1시간	45.5	194.2	13.9	0.7
6시간	109.0	1107.8	33.3	0.1
12시간	135.3	1638.2	40.5	0.1
24시간	159.8	3429.5	58.6	0.4

4.3 자료의 변환

EDA 기법이 유사 정규분포를 가정하기 때문에 강우와 같이 자료의 분포가 왜곡되어 있는 경우 특히 분산 및 왜곡도가 전통적 방법의 통계치보다 작게 산정되는 문제가 있음을 파악할 수 있었다. 따라서 확률강우량도 전통적 방법과 비교했을 때, EDA 기법을 적용한 경우가 작게 산정되는 것이다.

이러한 문제를 극복하기 위해서는 정규분포를 따르지 않는



(a) 확률밀도함수 (b) 확률강우량

그림 6. 전통적 방법(MOM)과 EDA 기법에 의한 방법의 확률밀도함수 비교

표 5. 전통적 방법과 EDA 기법의 적용에 따라 추정된 확률강우량 비교(1시간) (단위:mm)

재현기간	전통적 방법	EDA 기법
10	70.4	64.0
20	79.5	70.5
30	84.5	74.1
50	90.6	78.4
80	95.9	82.2
100	98.5	83.9

수문자료를 적절히 변환하여 정규분포를 따르도록 할 필요가 있다. 즉, 자료가 정규분포를 따르도록 변환하고, 변환된 자료에 대하여 EDA 기법을 적용하며, 마지막으로 그 결과를 다시 역변환하여 확률강우량의 추정에 이용하는 것이다. 본 연구에서는 자료 변환기법으로 Box and Cox(1964)에 의해 제안된 변환기법을 사용하여 자료의 정규화를 시도하였다. 다음 식은 Box-Cox 변환을 정의한 것이다.

$$y_t^{(\lambda)} = \begin{cases} \dfrac{x_t^{\lambda} - \lambda}{\lambda} & (\lambda \neq 0) \\ \ln x_t & (\lambda \neq 0) \end{cases} \tag{4}$$

여기서 y_t는 Box-Cox 변환을 통해 변환된 값이며 x_t는 원자료를 나타낸다. λ는 변환계수를 나타내며, 이 변환계수를 통하여 Box-Cox 변환이 일어난다. 변환계수 λ는 보통 -3.0에서 3.0 사이의 값을 범위로 한다. 변환된 자료의 역변환은 다음 식에 의해 수행된다.

$$x_t = (y_t \cdot \lambda + 1)^{\left(\frac{1}{\lambda}\right)} \tag{5}$$

본 연구에서는 변환계수 λ를 0.1단위로 변화시켜 가며 정규분포에 대한 적합도 검정(x^2-test, K-S test, PPCC test)을 수행하였다. 그 결과 검정을 모두 통과한 변환계수 λ의 값은 0.7에서 1.2구간의 값으로 파악되었다. 이 구간에 대해서는 추가로 변환계수 λ를 0.01로 더 세분하고 각각의 경우에 대해 추정된 표준편차와 왜곡도의 역변환 값이 원자료의 표준편차와 왜곡도를 얼마나 잘 재현하는지를 파악하였다. 그 결과 변환계수 λ가 1.12일 때 원자료의 왜곡도와 표준편차를 가장 잘 반영하는 것으로 확인되었다. 그림 7은 정규화변수 값의 변화에 따른 확률밀도함수를 나타낸 것이다.

표 6. 정규화 자료의 통계특성 역변환 (λ=1.12)

	원자료		변환된 자료 / EDA	
	전통적 방법 (MOM)	EDA	변환	역변환
평균	46.3	45.5	63.5	45.6
표준편차	18.1	13.9	22.0	18.1
왜곡도	1.4	0.2	0.2	1.2

4.4 변환된 자료에 대한 EDA 기법의 적용

EDA 기법은 유사 정규분포를 가정한 방법이므로 원자료가 왜곡된 분포일 경우 변환의 필요성이 있다. 이는 전 절에서 살펴본 바와 같다. 변환된 자료에 EDA 기법을 적용하면 평균, 표준편차 및 왜곡도 계수를 분위수를 이용하여 추정할 수 있고, 이를 역변환 하면 원자료에 대한 통계 추정치가 된다. 이 값은 물론 원자료에 EDA 기법을 적용하여 추정한 통계치와 다르다.

표 6은 EDA 기법을 적용하여 추정한 원자료 및 변환된 자료의 통계치를 나타낸 것이다. 변환된 자료의 통계치는 다시 역변환된 원자료의 통계치와 비교될 수 있게 하였다. 표

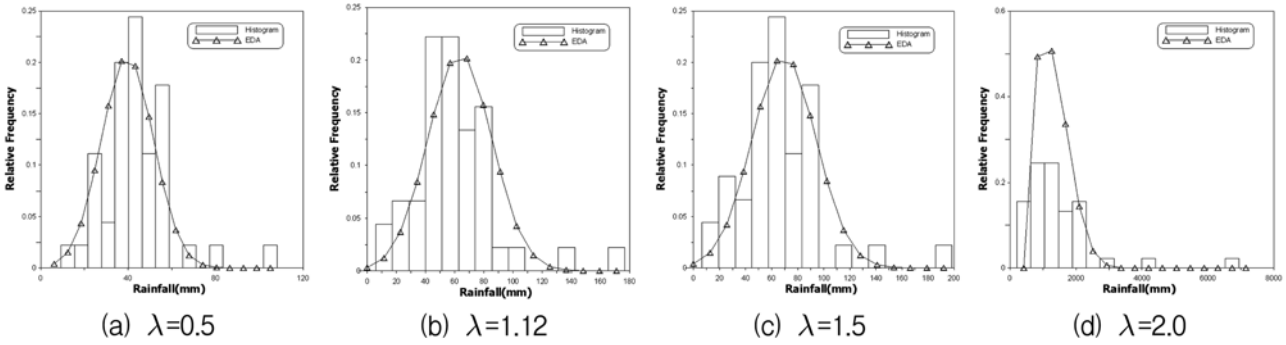


그림 7. λ값 변화에 따른 확률밀도함수 변화 (λ=1.12를 최적의 값으로 결정)

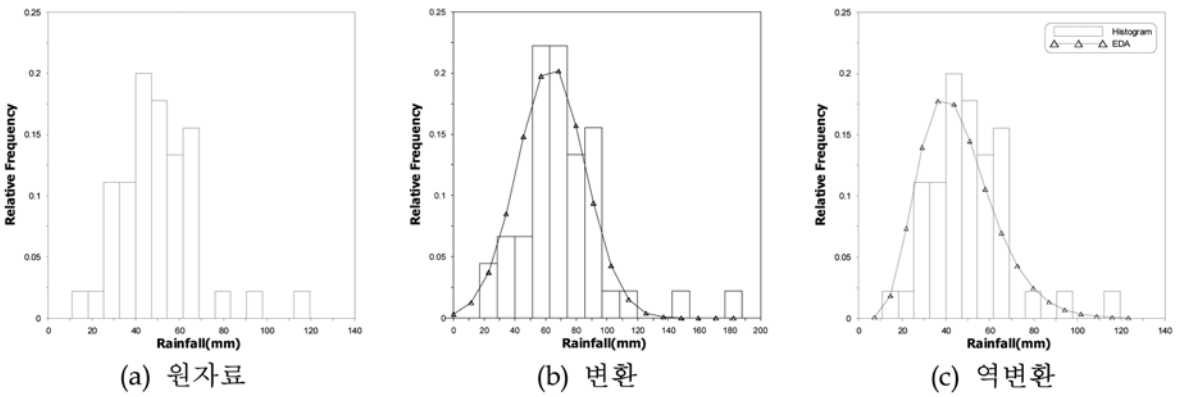


그림 8. 2변수 Gamma 분포의 변환역변환된 자료의 확률밀도함수

표 7. 1시간 지속시간 정규분포의 재현기간별 확률강우량(단위:mm)

재현기간	모멘트법	EDA 기법
10	60.4	60.0
20	70.4	69.8
30	79.5	78.9
50	84.5	83.9
80	90.6	90.1
100	96.0	95.5

6에 나타난 것과 같이 전통적 방법과 최종적으로 역변환된 통계치가 거의 비슷함을 알 수 있다. EDA 기법은 원자료에 EDA 기법을 적용한 경우와 달리, 변환-역변환 과정을 거치면서 특히 왜곡도가 원자료의 특성을 매우 잘 나타내는 것으로 확인되었다.

변환된 자료에 EDA 기법을 적용하여 추정한 통계치의 역변환 값은 확률밀도함수의 매개변수 추정에 이용된다. 본 연구에서는 2변수 Gamma 분포를 이용하였다. 추정된 2변수 Gamma 분포의 확률밀도함수는 그림 8과 같다. 여기서 그림 8(b)는 변환전 자료의 확률밀도함수, 그림 8(c)는 변환-역변환된 확률밀도함수를 나타낸 것이다. 표 7은 모멘트법에 근거하는 경우와 EDA 기법을 적용하는 경우에 추정한 확률강우량을 비교한 것이다. 최종적으로 추정된 통계치가 유사하므로(표 6), 확률강우량이 유사하게 추정되는 것은 당연하다.

5. 신규자료의 추가에 따른 확률강우량의 변동

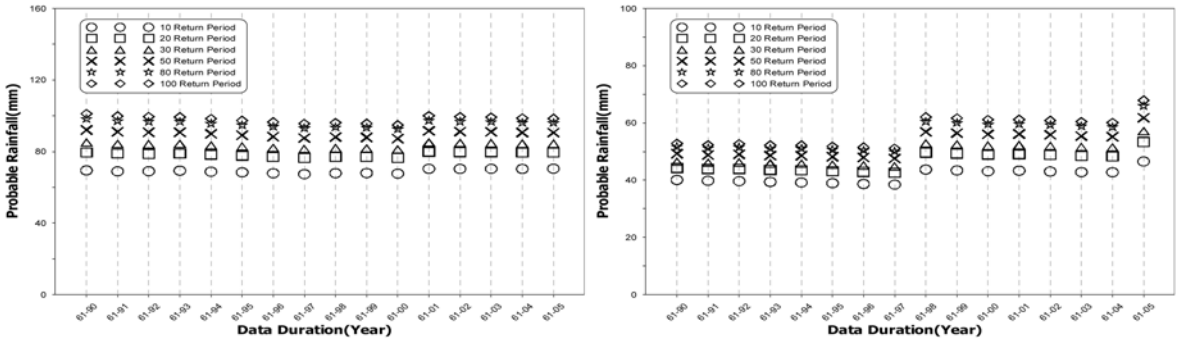
본 장에서는 매년 신규 강우자료의 추가에 따라 발생하는 확률강우량의 변동을 기존의 모멘트법에 근거한 경우와 본 연구에서 제시하는 EDA 기법을 적용하는 경우로 나누어 비교하였다. 일단 4장에서의 결과처럼 기존의 빈도해석 방법과 EDA 기법을 적용하는 경우 유사한 확률강우량이 추정된다는 것은 파악할 수 있었다. 그러나 EDA 기법을 적용하는 경우 EDA의 특징인 저항성을 근거로 특히 극치 규모의 이상 강우의 발생 시 보다 안정적인 확률강우량의 산정이 가능할 것으로 판단된다.

본 연구에서는 서울 및 포항 지점을 그 대상으로 하였으며, Ahn *et al.*(2003)의 연구에서 제시된 바와 같이 분석을 위한 최소 자료길이는 30년으로 하였다. 즉, 서울 및 포항 두 지점 모두 1961~1990년까지의 자료를 이용하여 빈도해석을 실시한 후, 1년씩 자료를 추가해 가며 신규자료의 영향을 판단하였다. 먼저 서울 및 포항지점의 연최대치 1시간 강우자료에 대한 전통적 빈도해석 결과는 각각 그림 9(a), (b)와 같다.

먼저 서울지점에 해당하는 그림 9(a)에서는 확률강우량의 변동이 아주 크지 않음을 보여 준다. 2001년에 약간의 상승이 확인되는데 이는 전체 평균치의 두 배 정도에 해당하는 신규자료(약 90 mm/hr)의 영향이다.

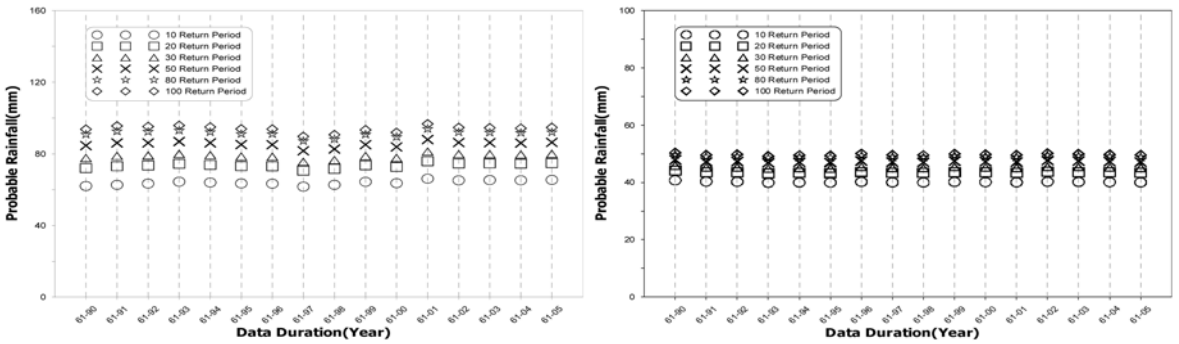
포항지점의 경우는 서울지점의 경우와 많이 다르다. 그림 9(b)에서 살펴볼 수 있는 것처럼 확률강우량의 급격한 변동을 확인할 수 있었다. 1998년과 2005년의 확률강우량 상승은 기록 평균치의 두 배 이상인 각각 75.5 mm/hr와 76.0 mm/hr의 영향이다. 즉, 평균치보다 월등히 큰 값의 추가로 인해 확률강우량이 크게 변동할 수 있음을 나타내는 결과로 해석할 수 있다.

동일한 자료에 EDA 기법을 적용하는 경우의 결과는 그림



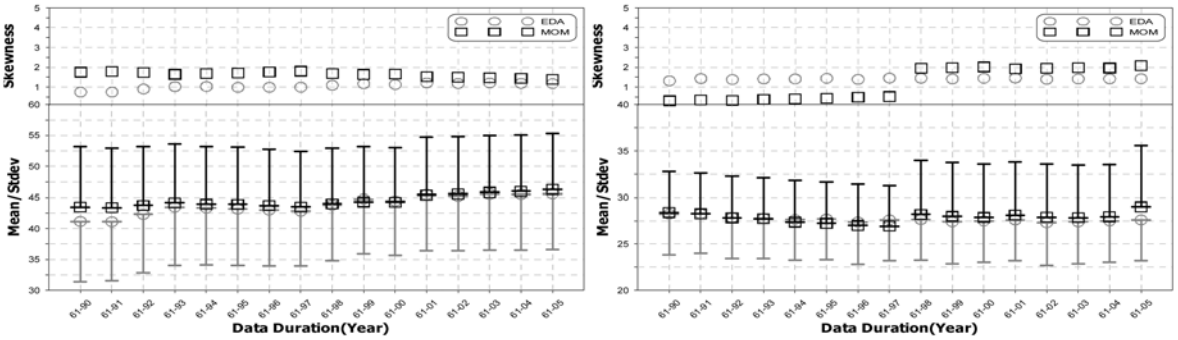
(a) 서울지점 (b) 포항지점

그림 9. 신규자료의 추가에 따른 재현기간별 확률강우량 변동(전통적 방법)



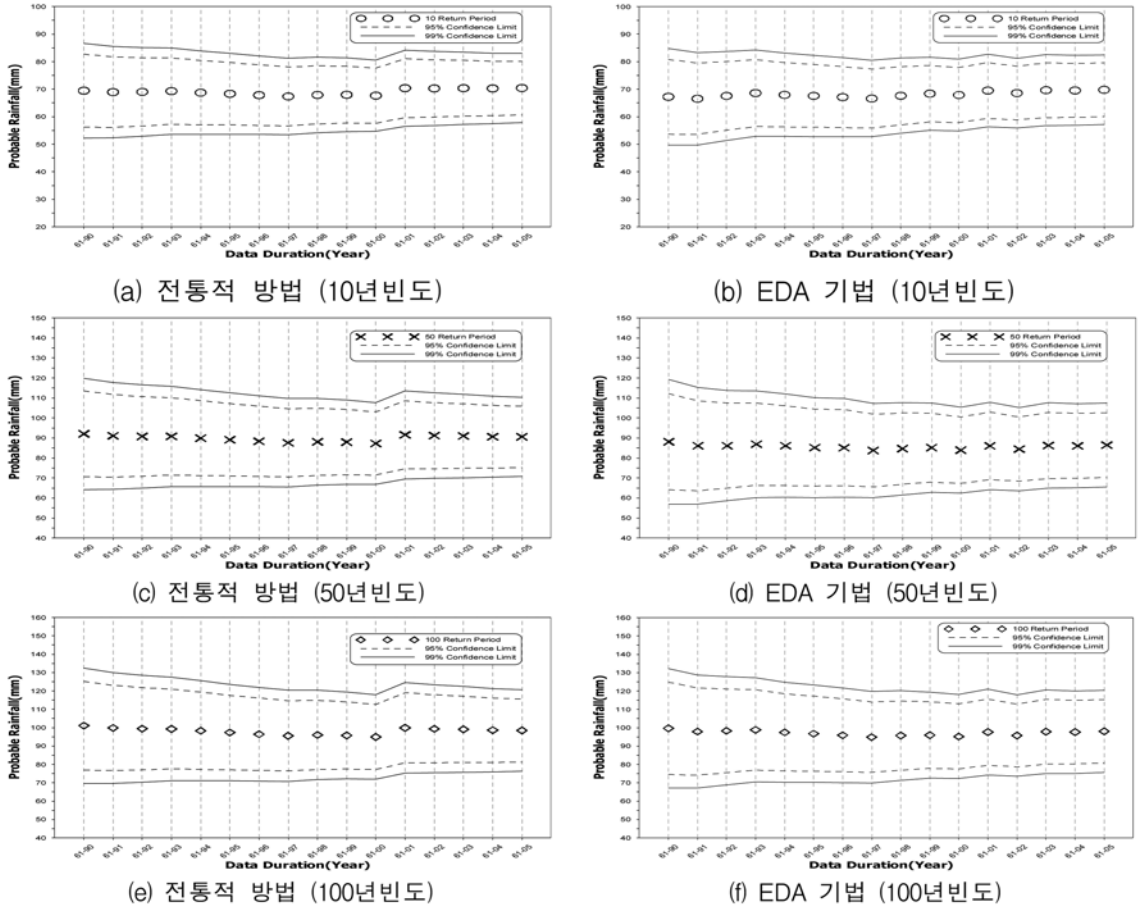
(a) 서울지점 (b) 포항지점

그림 10. 신규자료의 추가에 따른 재현기간별 확률강우량 변동(EDA 기법)



(a) 서울지점 (b) 포항지점

그림 11. 신규자료의 추가에 따른 평균, 표준편차, 왜곡도의 변화



(a) 전통적 방법 (10년빈도) (b) EDA 기법 (10년빈도)

(c) 전통적 방법 (50년빈도) (d) EDA 기법 (50년빈도)

(e) 전통적 방법 (100년빈도) (f) EDA 기법 (100년빈도)

그림 12. 신규자료의 추가에 따른 확률강우량 변동 (서울지점)

10과 같다. 이 결과는 EDA 기법 적용을 위한 자료의 변환-역변화 과정에 필요한 변환계수 λ를 매년 신규자료가 추가될 때마다 다시 추정한 경우에 해당하는 것이다. 본 논문에는 수록하지 않았으나 처음 30년간 자료의 변환-역변화 과정에서 결정된 변환계수 λ를 동일하게 적용한 경우에 있어서도 유사한 결과를 확인할 수 있었다(박현근, 2007). 이러한 결과는 물론 그림 11에 나타난 것과 같이 자료의 통계치가 다르게 추정되었기 때문이다. 전통적인 모멘트법에 근거한 통계치의 경우는 신규자료가 추가됨에 따라 크게 변하는 모습을 보여주지만 EDA 기법을 적용한 경우에 있어서는 상대적으로 안정된 값을 나타내는 것을 확인할 수 있었다. 특히 왜곡도의 경우에 아주 뚜렷한 차이를 확인할 수 있었다. 서울 지점과 포항지점을 비교하면 포항지점의 경우가 보다 큰 차이를 나타내고 있음을 확인할 수 있었다.

그림 12와 그림 13은 각각의 방법으로 추정된 확률강우량의 신뢰구간을 함께 도시한 것이다. 2변수 Gamma 분포의 확률강우량에 대한 신뢰구간은 다음 식과 같은 표준오차를 이용하여 결정할 수 있다(Rao and Hamed, 2000).

$$s_T^2 = \frac{\sigma^2}{N}\left[(1+K_T C_V)^2 + \frac{1}{2}\left(K_T + 2C_V\frac{\partial K_T}{C_s}\right)^2 (1+C_V^2)\right] \tag{6}$$

여기서, s_T는 표준오차이며, C_V는 변동계수, K_T는 빈도계수, C_S는 왜곡도계수를 뜻한다. $\partial K_T/\partial C_S$는 다음 식으로 계산된다.

$$\frac{\partial K_T}{\partial C_s} = \frac{-2}{C_S^2}\left[\left\{\frac{C_S}{6}\left(u-\frac{C_S}{6}\right)+1\right\}^3 - 1\right] + \frac{2}{C_S}\left[3\left\{\frac{C_S}{6}\left(u-\frac{C_S}{6}\right)+1\right\}^2\left(\frac{u}{6}-\frac{2C_S}{36}\right)\right] \tag{7}$$

여기서, σ^2은 2변수 Gamma 분포의 매개변수 α와 β를 이

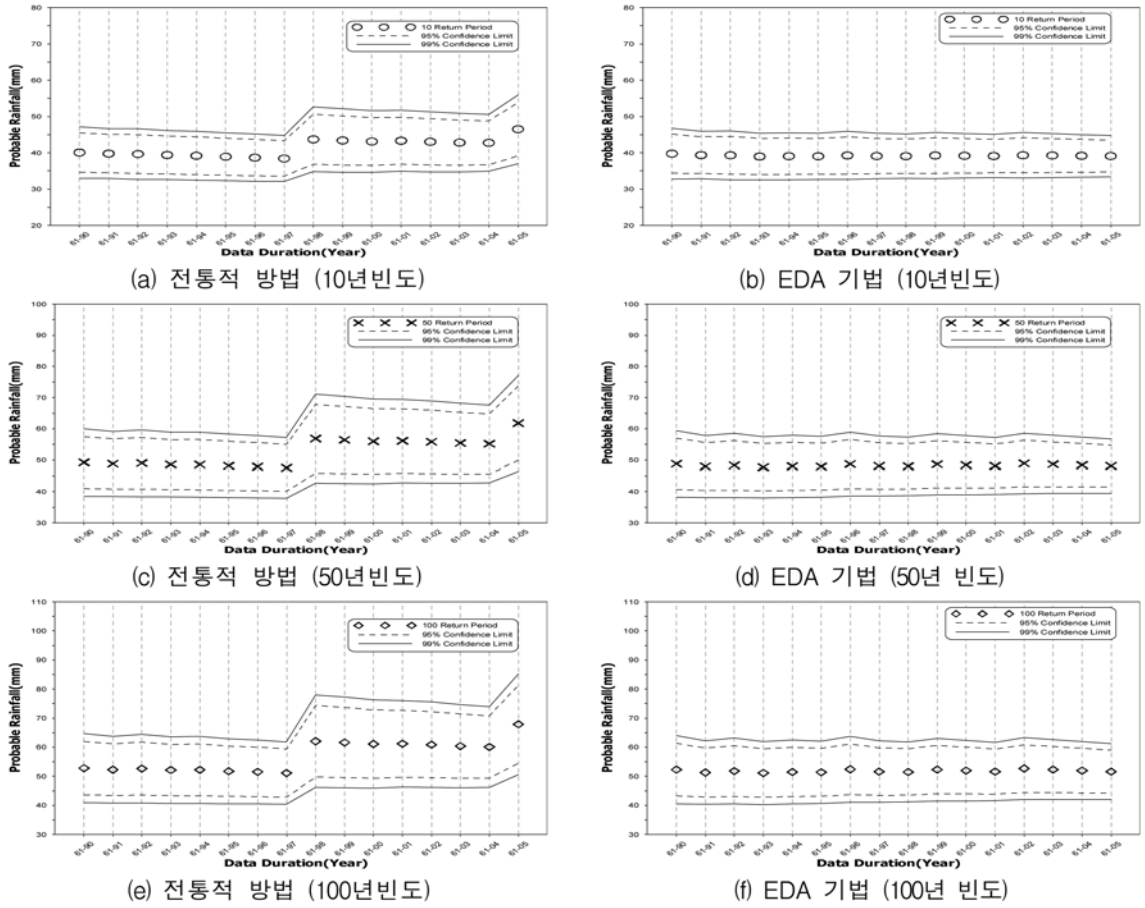


그림 13. 신규자료의 추가에 따른 확률강우량 변동 (포항지점)

용하여 다음과 같이 결정된다.

$$\sigma^2 = \alpha^2 \beta \tag{8}$$

따라서 유의수준 95% 및 99% 신뢰구간은 각각 다음과 같이 결정된다.

$$\hat{x} - 1.96 S_T < \hat{x} < \hat{x} + 1.96 S_T \tag{9}$$

$$\hat{x} - 2.54 S_T < \hat{x} < \hat{x} + 2.54 S_T \tag{10}$$

먼저 그림 12(a)~(f)는 서울지점에 대한 결과를 표시한 것이다. 그림 12(a), (c), (e)는 전통적인 빈도해석의 결과에 해당하며, 그림 12(b), (d), (f)는 EDA 기법을 적용한 경우에 해당한다. 그림에서 살펴볼 수 있는 것처럼 EDA 기법을 적용한 경우가 보다 안정적인 확률강우량을 추정해 줌을 확인할 수 있었다. 이러한 결과는 포함지점의 경우에 보다 뚜렷하게 확인할 수 있었다(그림 13). 확률강우량의 신뢰구간 또한 유사한 경향임을 확인할 수 있었다. 전통적인 방법의 빈도해석 결과에서는 극치규모의 강우사상이 추가되는 경우 확률강우량의 신뢰구간 역시 크게 넓어지는 경향을 보이지만, EDA 기법을 적용할 경우에는 그 변화의 폭이 훨씬 적을 뿐만 아니라 자료 기간의 증가에 따라 신뢰구간이 축소되어 가는 안정적인 경향을 뚜렷하게 보여주고 있다. 이러한 결과는 물론 EDA 기법의 특성인 저항성에서 그 원인을 찾을 수 있다.

6. 결 론

본 연구에서는 이상치의 포함여부에 따라 큰 변동폭을 보일 수밖에 없는 전통적 확률강우량 산정방법의 문제점을 보완하기 위한 방안으로 탐색적 자료분석(exploratory data analysis: EDA) 기법의 적용성을 검토하였다. EDA 기법에서 사용하는 자료의 특성화는 자료의 구조에 대한 것으로 전통적인 방법이 모멘트법에 근거하여 자료자체의 크기를 고려하는 것과 대비된다. 본 연구는 서울 및 포항지점에 적용하였으며, 그 결과를 정리하면 다음과 같다.

1. 원자료에 EDA 기법을 바로 적용했을 경우 추정된 확률강우량은 전통적 방법에 비하여 상당히 작게 산정된다. 이는 원자료가 오른쪽으로 왜곡된 분포형을 나타내기 때문이며, 따라서 재현기간이 큰 확률강우량이 과소 추정될 여지가 크다. 이러한 문제점은 자료가 정규분포형을 따르도록 변환한 후 EDA 기법을 적용하고, 추정된 통계치를 역변환 함으로서 극복할 수 있었다. 결과적으로, 자료변환 및 EDA 기법의 적용을 통해 유도된 빈도해석 결과는 전통적 빈도해석 결과와 매우 유사한 것을 파악할 수 있었다.
2. 신규자료의 추가에 따른 확률강우량의 변동은 EDA 기법을 적용함으로서 크게 완화할 수 있었다. 이는 EDA 기법이 개개 관측자료가 아닌 자료의 구조를 고려함으로서 생기는 저항성 때문인 것으로 판단된다.

감사의 글

본 연구는 국토해양부가 출연하고 한국건설교통기술평가원에서 위탁시행 한 2003년도 건설핵심기술연구개발사업(03산

학연C01-01)에 의한 도시홍수재해관리기술연구사업단의 연구성과입니다.

참고문헌

국립방재연구소(2002) **2002 태풍루사 피해 현장조사 보고서**. 행정자치부, 국립방재연구소.

박상덕(2002) 태풍 루사로 인한 홍수특성과 대책, **한국수자원학회논문집**, 한국수자원학회, 제35권 제6호, pp. 36-47.

박현근(2007) **확률강우량 산정을 위한 EDA 기법의 적용**. 석사학위논문, 고려대학교.

백운붕, 허명회(1987) **EDA-탐색적 데이터분석**. 박영사.

안재현, 김태웅, 유철상, 윤용남(2000) 자료기간에 따른 우리나라 확률강우량의 변화 분석, **한국수자원학회논문집**, 한국수자원학회, 제33권 제제5호, pp. 569-580.

유철상, 정성인, 윤용남(2007) 확률강우량의 정상성 판단: 2. 새로운 방법의 제안, **한국방재학회논문집**, 한국방재학회, 제7권 제5호, pp. 99-107.

윤용남(1998) **공업수문학**. 청문각.

정성인, 유철상, 윤용남(2007) 확률강우량의 정상성 판단: 1. 기존 방법의 적용, **한국방재학회논문집**, 한국방재학회, 제7권 제5호, pp. 89-98.

정종호, 윤용남(2003) **수자원 설계실무.** 구미서관.

차은정, 최영진(2000) 한반도 여름철 집중호우의 시간 · 공간 변동 특성, **한국수자원학회학술기사**, 한국수자원학회, 제33권 제4호, pp 47-56.

한국건설기술연구원(2000) **수자원계획의 최적화 연구**(Ⅳ) : **기후변화에 따른 수자원 계획의 영향 평가**. 건설교통부, 한국수자원공사, pp. 344-347.

한화진, 안소은, 최은진, 한기주, 이정택, 김해동, 손요한, 박용하, 조광우, 윤정호, 이은애, 김승만(2005) **기후변화 영향평가 및 적응시스템 구축 Ⅰ**. 한국환경정책 · 평가연구원, pp. 219.

허명회, 이태림(1993) **탐색적 자료분석.** 한국방송통신대학교 출판부.

Ahn, J., Kim, T., Yoo, C., and Yoon, Y. (2003) On the variation of frequency-based rainfall amounts : a case study for evaluating recent extreme rainfall in korea, *Stochastic Environmental Research and Risk Assessment*, Vol. 17, pp. 217-227.

Box, G. and Cox, D. (1964) An analysis of transformations (with discussion), *Journal of the Royal Statistical Society* (B), Vol. 26, pp. 211-252.

Coles, S., Pericchi, L.R., and Sisson, S. (2003) A fully probabilistic approach to extreme rainfall modeling, *Journal of Hydrology*, Vol. 273, pp. 35-50.

Rao, A. and Hamed, K. (2000) *Flood Frequency Analysis*. CRC.

Tukey, J. (1977) Exploratory data analysis, Addison-Wesley Pub. Co.

Wang, Q.J. (1997) Using Higher probability weighted moments for flood frequency analysis, *Journal of Hydrology*, Vol. 194, pp. 95-106.

(접수일: 2008.1.28/심사일: 2008.3.14/심사완료일: 2008.5.30)