응용통계연구 제 **20**권 *1*호, *2007*년, *pp. 1–9*

일일 최고기온의 변화에 대한 추정*

고왕경[1)]

요 약

한국의 네 개 도시(서울, 대구, 춘천, 영천)의 일일 최고기온을 모형화하여, 이에 적합한 분포를 제안하고 분포의 적합성을 여러 가지 방법에 의하여 검토하였다. 제안된 분포는 극단값 분포의 일종이며, 적합성 검토는 카이제곱 적합도 검정, Q-Q plot, 확률그림과 5000번의 모의실험을 통하여 허용한계를 구하였다. 그 결과 제안된 극단값 분포(Extreme Value Distribution)가 일일 최고기온을 잘 설명하고 있음을 확인할 수 있었다. 논문에서 나타난 실제 데이터의 그림은 서울의 1월과 6월을 중심으로 하였고, 대상지역의 2006년과 100년 후 2105년의 평균기온과, 제안된 극단값 분포에 의해 95% 신뢰구간하에서 일일 최고기온의 평균 상한값을 예측하였다.

주요용어: 일일 최고기온, 극단값 분포, Q-Q plot, 확률그림, 허용한계.

1. 서론

여러 가지 보고에 따르면 지구 온난화 현상에 의해 지구가 덥혀지고 있다고 한다. 이런 현상이 우리나라에도 영향력이 미치는지를 예측하는 것은 매우 중요한 일이다.

따라서 이 논문은 측정된 일일 최고기온 자료를 이용하여 기온의 변화를 추정하고 이들에 적합한 분포를 제안하고, 제안된 함수를 이용하여 일일 최고기온을 추정함을 목적으로 한다. 분석대상이 되는 일일 최고기온(daily maxima temperature)은 시간의 경과에 따라 변동하는 변수이지만 시간적 변동을 고려 대상에 넣지 않고 각 월별로 일차원적인 분석을 중심으로 하였다.

분석대상 지역은 우리나라 네 지역(서울, 대구, 춘천, 영천)이다. 각 지역을 월별로 일일 최고기온을 분석하여 이에 적합한 극단값 분포를 제안하고, 그 적합도를 여러 가지 통계적 검정을 통하여 수행한다. 또한 제안한 분포를 이용하여 100년 후 2105년의 일일 평균 최고기온과 95% 신뢰구간에서 상한값을 예측한다.

분석에 사용된 일일 최고기온은 온라인상에서 얻은 자료로서, 서울과 대구 지역은 1961년부터 2005년까지 45년간이며, 춘천지역은 1966년부터 2005년 동안의 40년간, 영천지역은 1973년부터 2005년까지 33년의 일일 최고기온 기록에 의한 것이다. 또한 논문상에 표현된 여러 가지 그림들은 서울 지역의 1월과 6월을 중심으로 한다. 분석 결과에 대한 해석은 통계적인 해석을 위주로 하였고 기상학적인 해석은 가미하지 않았음을 밝혀두는 바이다. 또

* 본 논문은 안양대학교 안식년 기간 중 연구되었음.

1) (430-714) 경기도 안양시 만안구 안양5동 708-113, 안양대학교 정보통계학과, 교수
E-mail: wkko@aycc.anyang.ac.kr.

한 이 논문에서 나타나는 모든 분석의 결과는 기온 변화에 큰 순환(large cycle)이 없다는 가정 하에서 출발하며, 분석에 사용된 모든 프로그램은 Maple 언어로 작성하였다.

제 2절에서는 일일 최고기온이 어떤 추세를 갖고 있는지를 추세선을 이용하여 대상 지역에서 대상 기간의 기온 증가(감소)를 구하고, 일일 최고기온에 적합시킨 극단값 분포를 제안하고 실제 데이터와의 관계를 여러 가지 방법(카이제곱 적합도 검정, Q-Q plot, 확률그림, 허용한계)에 의해 검토하였다. 제 3절에서는 제안된 극단값 분포 하에서 2105년의 평균 일일 최고 기온과 95% 신뢰구간 하에서 평균 상한치를 추정하였으며, 제 4절에서는 우리의 관심사에 대한 결론을 유도하고 추후 과제를 논의하겠다.

2. 분포와 통계적 분석

2.1. 일일 최고기온의 추세

그림 2.1은 서울지역의 지난 45년 동안 1월과 6월의 일일 최고기온의 산포와 추세선을 나타낸 것이다. 그림에서 확인하는 바와 같이 일일 최고기온의 추세선은 완만한 상승을 나타내어 매년 일일 최고기온이 상승하고 있음을 알 수 있다.

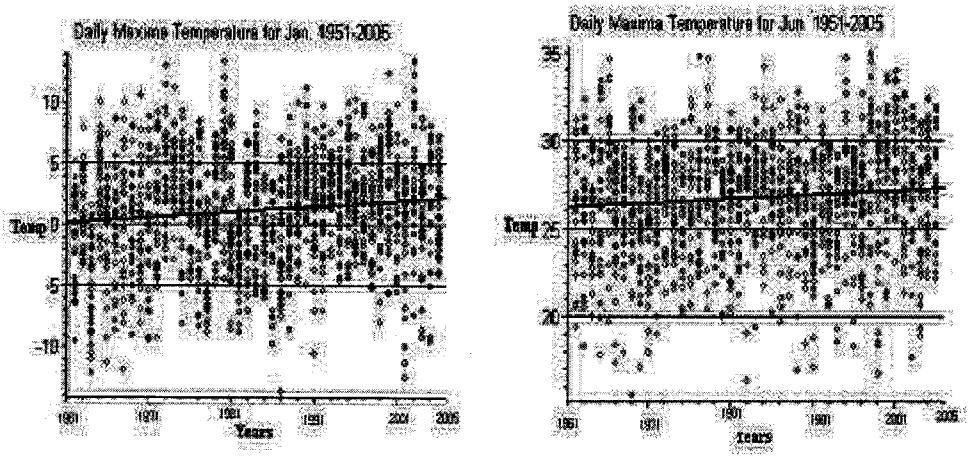

그림 2.1: 45년간의 일일 최고기온과 추세선

또한 그림 2.2는 일일 최고기온을 시계열 자료 형태로 표현한 것이며, 굵게 표현된 직선은 시계열 자료의 추세선을 의미한다. 추세선은 앞의 그림 2.1과 마찬가지로 완만한 상승을 보이고 있다. 단, 가로축은 날짜 순으로 표시한 것으로서, 1961년 1월 1일이나 6월 1일은 1로 나타내며 2005년 1월 31은 1395로, 2005년 6월 30일은 1350으로 표시되었다.

추세선의 분석에 의하면 각 지역의 측정된 기간에 월별 일일 최고기온 평균의 증가(또는 감소)는 표 2.1과 같다. 표에서 서울 지역은 45년 동안 매월평균 1.14°C만큼 기온상승이 있었고, 대구 지역은 1.47°C 상승하였으며, 춘천 지역은 40년 동안 1.55°C가 상승하였고 영

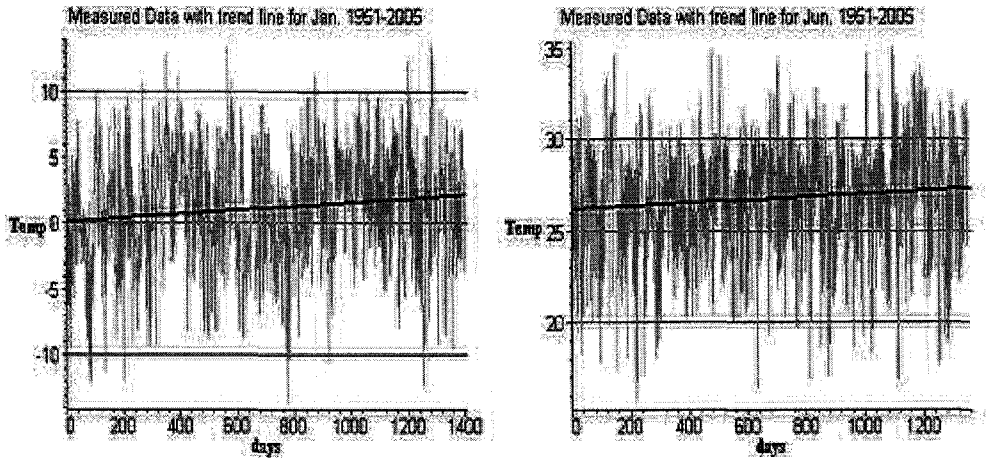

그림 2.2: 45년간의 실제 데이터와 추세선

천지역은 33년 동안 0.93°C의 상승을 보여주고 있다. 특이한 것은 네 지역 모두 측정 기간 내의 평균 상승률이 5월에서부터 10월까지는 하강(특히 8월)하거나 1°C보다 작은 상승을 보인 반면 나머지 기간은 1°C 이상의 상승을 보이고 있다는 점과 춘천지역이 측정기간이 짧은데도 불구하고 년 평균 기온상승률(0.03875°C)이 타 지역에 비하여 상대적으로 높은 것으로 나타났다.

서울의 1월의 추세를 분석해 보면 매년 평균적으로 0.0448°C씩 증가하여 2105년에는 일일 평균 최고기온은 기준년도보다 6.4996°C 만큼 상승하여 6.6731°C가 되며, 6월은 매년 평균적으로 0.0236°C가 증가하여 2105년에는 3.4256°C의 증가가 생겨 일일 평균 최고기온은 29.8184°C가 되는 것으로 나타났다.

2.2. 극단값 분포

일일 최고기온은 시간의 변화에 따른 변수이다. 그러나 여기서는 시간 변수를 고려하지 않고 일일 최고기온에 적합한 분포(Beirlant 등(2004), Galambos 등(1993))를 모형화하겠다.

표 2.1: 월별 일일 최고기온의 변화량(단위:°C)

지역	기간	1	2	3	4	5	6	7	8	9	10	11	12	평균
서울	45년	2.02	3.03	2.53	1.69	0.05	1.12	0.36	-0.22	0.69	0.03	1.05	1.35	1.14
대구	45년	2.53	3.49	2.56	3.42	0.72	0.93	-0.08	-1.47	0.83	0.83	1.99	1.92	1.47
춘천	40년	1.47	3.26	3.12	1.74	0.23	1.42	0.82	0.60	1.47	0.69	1.65	2.16	1.55
영천	33년	1.17	2.23	1.61	2.13	0.45	0.72	0.10	-0.70	0.93	-0.08	1.78	0.83	0.93

일일 최고기온을 확률변수X라 하면, 확률변수X는 두 개의 모수 즉, 평균 μ와 표준편차 σ를 갖는 다음과 같은 극단값 분포함수를 가정하겠다.

$$\begin{aligned} F(x) &= P(X \le x) \\ &= 1 - \exp\left\{-\exp\left(-\psi(1) - \sqrt{\psi'(1)}\frac{(y-\mu)}{\sigma}\right)\right\}. \end{aligned}$$

이때, $y = -\ln(x+m)$이며, $\psi(1) = 0.5772$, $\psi'(1) = \pi^2/6$인 Euler 상수이다.

이 분포의 모양은 m과 σ의 값에 민감한 반응을 보이는 함수이다. 따라서 이 논문에서는 $m = \min \sum_{i=1}^{k}\{(o_i - e_i)^2\}/e_i$ (단, o_i는 관측치이고 e_i는 이론치)의 값을 사용하였다.

2.3. 여러 가지 적합도 검정

앞에서 제안한 극단값 분포 $F(x)$가 분석 대상인 일일 최고기온를 설명하는데 적절한지를 여러 가지 통계적 방법을 통하여 검토해 보기로 하자.

첫 번째는 카이제곱 통계량을 이용한 적합도 검정이다. 주어진 일일 최고기온을 20개의 균일 구간으로 구분하여 각 구간에 해당하는 도수와 함수 $f(x)$를 적용시켰을 때 각 구간에 해당하는 이론치 간에 차이가 있는지를 검정하기 위하여 카이제곱 적합도 검정통계량을 이용하였다. 표 2.2는 카이제곱 통계량 값과 이에 해당하는 $p-$값을 지역별, 월별로 나타낸 것이다.

표 2.2: 카이제곱 통계량 값과 $p-$값(단위:°C)

지역		월					
		1	2	3	4	5	6
서울	통계량	13.6489	29.7575	22.4816	18.6006	32.3055	24.6873
	$p-$값	.8038	.550	.2609	.4827	.0289	.1711
대구	통계량	43.3000	41.6944	33.3596	22.5274	34.4667	41.8888
	$p-$값	.0012*	.0019*	.0218*	.2588	.0162*	.0018*
춘천	통계량	28.5327	25.6403	17.3001	16.2385	29.8729	30.8369
	$p-$값	.0737	.1405	.5695	.6413	.0534	.0421*
영천	통계량	19.4886	47.3795	35.8465	20.6251	24.0752	46.6195
	$p-$값	.4259	.0003*	.0110*	.3579	.1933	.0004*
지역		월					
		7	8	9	10	11	12
서울	통계량	61.6555	22.1069	22.0365	26.7339	37.0464	16.6643
	$p-$값	.0000*	.2790	.2824	.1110	.0078*	.6126
대구	통계량	69.6501	19.0457	13.8454	31.5056	11.0091	24.9795
	$p-$값	.0000*	.4539	.7927	.0355*	.9235	.1612
춘천	통계량	38.8830	17.6542	21.4450	15.0870	14.8593	10.7362
	$p-$값	.0046*	.5456	.3127	.7171	.7315	.9324
영천	통계량	82.2287	36.5333	14.3176	12.2588	16.8022	14.6927
	$p-$값	.0000*	.0091*	.7649	.8743	.6033	.7419

표에서 서울 지역과 춘천 지역은 비교적 높은 정도의 적합성을 보이고 있으나 대구 및 영천 지역은 그 적합도가 떨어짐을 알 수 있다. 이는 적절한 m을 선택하는데 문제가 있음을 나타내는 것이다.

그림 2.3은 서울의 1월과 6월의 관측된 일일 최고기온의 막대그래프와 제안된 극단값 분포에 적합시킨 확률 함수의 그림이다. 그림에서 보는 바와 같이 제안된 확률함수는 실제 데이터에 매우 잘 적합됨을 알 수 있다.

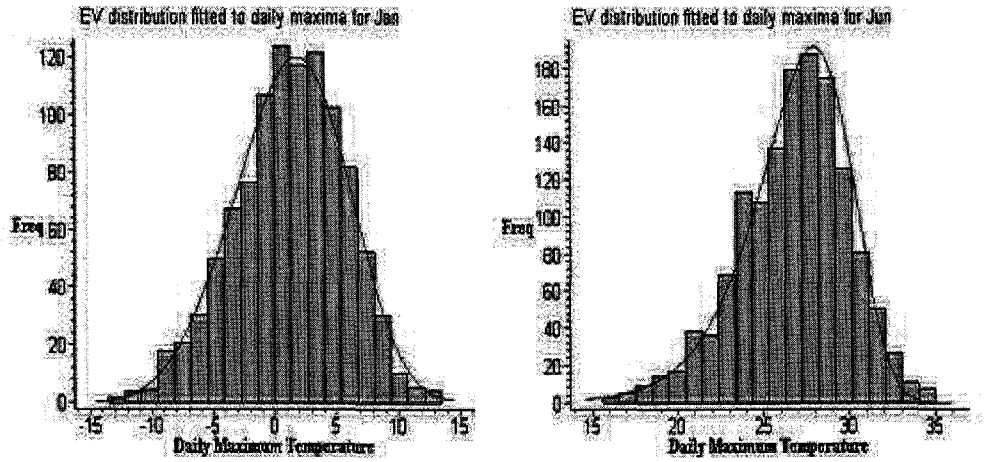

그림 2.3: 막대그래프와 극단값 분포

두 번째는 제안된 극단값 분포가 관측된 자료에 잘 적합되는지를 점검하기 위하여 Q-Q 그림(참조: 그림 2.4)을 그려 확인하는 것이다. 이 때 사용한 확률표본은 실제 데이터와 같은 크기의 임의표본을 사용하였으며, 모의실험을 할 때 극단값 분포의 분위수(quantile) 함수는

$$Q(p) = \exp\left(\frac{\sigma}{\sqrt{\psi'(1)}}(\psi(1) + \ln(-\ln(1-p))) - \mu\right) - m$$

에 의하여 구하여 졌다. 이 때 p는 (0, 1)사이의 난수이며, 또한 모수의 추정치는 실제 데이터에서 구한 값이다.

그림 2.4에서 가로축이 관측된 실제 데이터이고, 세로축은 분위수 함수에서 구한 임의표본인 Q-Q plot이다. 그림에서 보는 바와 같이 직선에 거의 가까운 산포를 보이고 있다. 따라서 관측치와 제안된 분포에서의 난수 사이에 그 적합도가 뛰어남을 알 수 있다.

세 번째 방법으로 적합도를 관찰하기 위하여 또 다른 분포 적합검증 방법인 확률그림(Probability Plot)(Gnanadesikan, 1997)을 그리면 그림 2.5와 같다.

이 그림에서 실선부분은 제안된 분포함수이고 점선 부분은 실제데이터의 누적분포함수의 그래프이다. 이 그림에서 확인하는 바와 같이 두 개의 곡선, 즉 두 개의 분포함수는 같은 형태임을 확인할 수 있다.

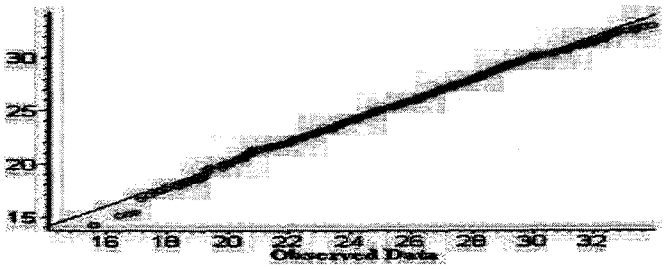

그림 2.4: 모의시험자료와 서울 6월의 Q-Q 그림

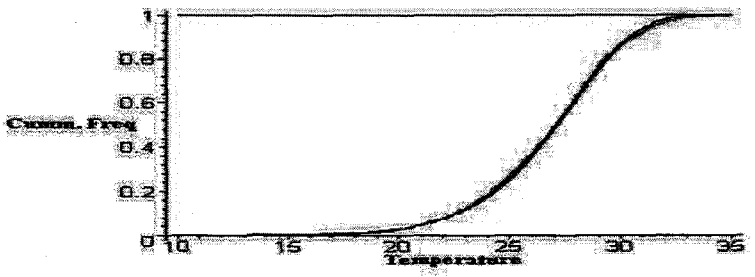

그림 2.5: 서울 6월의 확률 그림

네 번째 방법은 허용한계(Kendall과 Stuart, 1967)를 구하여 적합여부를 판정하는 방법이다. 제안한 분포에서 30개의 난수를 만들어 일일 최고기온 평균의 95% 신뢰구간을 만들고, 이 구간이 서울 6월 일일 최고기온 평균의 95% 신뢰구간인 $16.3358 \leq \mu \leq 37.4180$의 범위를 벗어나는지 아닌지를 5000번의 시뮬레이션을 시행하였다. 그 결과 오직 한 번 그 범위를 벗어남을 확인하였다.

이는 주어진 표본으로부터 모집단의 95%가 16.3358과 37.4180사이에 포함된다고 할 수 있는데, 이때 이와 같은 방법으로 5000번 반복한다면 4999번은 일일 최고온도 평균이 신뢰구간에 포함된다는 것이다(이재창, 1983). 따라서 이 값은 너무 작기 때문에 극단값 분포는 실제 데이터에 적합하다는 귀무가설을 기각할 수가 없음을 알 수 있다.

이와 같이 여러 가지 분포 적합도 검증 방법에 의하여 일일 최고기온의 확률분포는 앞에서 제안한 분포함수 $F(x)$임을 확인할 수 있다.

3. 제안된 분포에 따른 일일 최고기온 추정

앞에서 제안한 분포 함수$F(x)$를 적용하여 대상 지역의 2006년도와 2105년의 일일 최고기온의 평균과 95% 신뢰구간에서 일일 최고기온의 상한값(표에서 max로 표시)을 표 3.1에 구하였다.

월별 일일 최고기온의 변화량에서 보는 것과 같이 5월에서 10월까지는 낮은 정도로 상승(하강)하나 나머지 달은 약간 큰 폭으로 변동하고 있음을 알 수 있다. 표 중에서 과추정치(over-estimate, 예로서 춘천 지역의 7, 8월)로 생각되는 값들이 보임은 m의 값이 최적이 아님을 나타낸다. 그런데도 불구하고 대상지역에서 전반적으로 기온이 상승하고 있음을 확인할 수 있다.

표 3.1: 네 지역의 월 평균 년 최고기온 변화(단위:°C)

지역	년	추정	월											
			1	2	3	4	5	6	7	8	9	10	11	12
서울	2006년	평균	2.25	5.52	11.21	18.57	23.10	27.36	28.92	29.36	25.96	19.62	11.88	4.63
		max	3.34	8.85	17.36	28.58	33.77	37.42	40.76	38.86	36.28	23.91	16.27	7.11
	2015년	평균	6.66	12.42	16.58	22.62	23.45	29.82	29.88	28.75	27.26	19.32	13.70	7.35
		max	9.88	19.94	25.69	34.81	34.29	40.79	42.11	38.04	38.09	23.55	18.76	11.28
대구	2006년	평균	6.26	9.15	14.25	21.74	25.34	28.47	30.31	30.19	26.84	21.86	15.14	8.64
		max	8.67	16.06	23.06	34.87	34.56	26.45	38.84	40.30	44.86	29.69	22.27	13.25
	2015년	평균	11.79	17.04	20.18	29.56	27.15	30.55	30.37	26.77	28.43	23.39	19.08	12.65
		max	16.33	29.89	32.67	37.42	37.03	39.12	38.91	35.75	47.51	31.77	28.07	19.42
춘천	2006년	평균	1.87	6.02	12.35	19.84	24.04	28.13	29.57	29.74	25.90	19.57	11.49	4.54
		max	2.75	9.72	20.92	31.69	30.47	35.98	42.20	39.86	36.89	24.22	17.52	7.28
	2015년	평균	5.43	14.41	20.56	24.59	24.87	31.76	31.81	31.03	29.25	20.86	14.93	9.60
		max	8.01	23.28	34.84	39.28	31.52	40.63	45.41	41.59	41.66	25.81	22.74	15.39
영천	2006년	평균	5.48	8.53	13.58	20.91	24.64	27.65	29.72	29.67	25.97	21.07	14.87	7.86
		max	8.67	14.67	22.11	35.39	33.94	37.42	41.12	43.56	44.14	27.43	22.80	12.80
	2015년	평균	8.90	15.57	18.98	27.88	26.41	29.84	30.43	27.23	25.80	20.27	19.42	9.95
		max	14.10	26.76	30.90	37.18	36.38	40.38	42.11	39.98	43.85	26.40	29.78	16.20

4. 결론 및 제언

시간에 종속되어 있는 데이터의 특성에도 불구하고 일일 최고기온에 대한 분포로서 극단값 분포는 매우 만족스러울 정도로 적합성이 좋았다. 카이제곱 적합도 검정에서 대체로 만족스러운 결과를 얻었으며, Q-Q plot이나 확률그림으로의 판정도 매우 만족스러웠다. 또한 5000번의 시뮬레이션의 결과는 매우 만족스러운 결과를 얻었다.

극단값 분포에 따른 일일 최고기온의 추정에서 약간의 특이점들이 보이지만 비교적 만족스러울 정도의 예측을 할 수 있었다. 하여튼 제안된 극단값 분포를 이용한 추정은 매년 기온이 빠른 속도로 증가함을 보여준다. 추후 과제로서 m의 추정과, 이 분포를 기초로 하여 큰 순환이 존재하는 경우에 시계열 분석을 이용하여 순환변동 효과를 제거한 후에 일일 최고기온을 추정하는 것이다.

감사의 말씀

분포함수 유도에 도움을 주신 University of the Free State, South Africa 교수님이신 De Waal 교수님에게 감사의 말씀을 전합니다.

참고문헌

이재창 (1983). <통계학연습-통계수치표와 사용법->, 박영사, 서울.

Beirlant, J., Goegebeur, Y., Segers, J. and Teugels, J. (2004). *Statistics of Extremes: Theory and application*, John Wiley & Sons, England.

Galambos, J., Lechner, J., Simiu, E. and Macri, N. (1993). *Extreme Value Theory and Applications*, Kluwer Academic Pub., London.

Gnanadesikan, R. (1997). *Methods for Statistical Data Analysis of Multivariate Observations*, 2nd ed., John Wiley & Sons, New York.

Kendall, M. G. and Stuart, A. (1967). *The Advanced Theory of Statistics*, **2**, 2nd ed., Hafner Pub. co., New York.

[2006년 7월 접수, 2006년 8월 채택]

Estimation for the Change of Daily Maxima Temperature*

Wang Kyung Ko[1)]

ABSTRACT

This investigation on the change of the daily maxima temperature in Seoul, Daegu, Chunchen, Youngchen was triggered by news items such as the earth is getting warmer and a recent news item that said that Korea is getting warmer due to this climatic change. A statistical analysis on the daily maxima for June over this period in Seoul revealed a positive trend of 1.1190 centigrade over the 45 years, a change of 0.0249 degrees annually. Due to the large variation on these maximum temperatures, one can raise the question on the significance of this increase. To check the goodness of fit of the proposed extreme value model, we shown a Q-Q plot of the observed quantiles against the simulated quantiles and a probability plot. And we calculated statistics each month and a tolerance limit. This is tested through simulating a large number of similar datasets from an Extreme Value distribution which described the observed data very well. Only 0.02% of the simulated datasets showed an increase of this degrees or larger, meaning that the probability is very low for such an event to occur.

Keywords: Daily maxima temperature, extreme value distribution, Q-Q plot, probability plot, tolerance limit.

* This work was supported by Anyang University.

1) Professor, Department of Information and Statistics, Anyang University, Anyang 430-714, Korea
E-mail: wkko@aycc.anyang.ac.kr