공간통계모형에서 Box-Cox 변환에 대한 영향력 분석연구

이 진희[1] 신 기일[2]

요약

시계열 자료의 분석에서 분산이 일정하지 않을 경우 이에 대한 해결방법으로 변환이 사용된다. 그러나 이러한 변환은 분산을 안정화시킴으로서 추정 및 검정에 타당성을 주는 반면 새로운 편의를 생성하거나(Granger & Newbold,1976) 모형을 복잡하게 만듦으로써 해석의 어려움도 수반한다. 신과 강(2001)은 평균이 크고 그에 비해 분산이 작을 경우 Box-Cox 멱 변환이 시계열 자료에 대하여 별 영향을 미치지 않음을 연구하였다. 본 논문은 이에 대한 확장으로 공간자료에서도 이 이론이 성립함을 밝혔다.

주요용어 : Box-Cox 멱 변환, Variogram, 공간통계

1. 서론

Box-Cox 멱 변환은 분산이 일정하지 않을 경우 분산 안정화를 위하여 사용된다. 많은 통계적 분석, 즉 예측을 필요로 하는 모형을 설정함에 있어 우리는 분산이 일정 하다는 가정을 하게된다. 변환에 대한 기존의 방법들을 살펴보면 먼저 Jereny(1986)는 선형회귀에서 최적의 추정량을 찾기 위해 Box-Cox 멱 변환을 적합 시켰다. 그리고 그 결과에 대한 평균을 재 변환한 바 선형회귀에서의 smearing 추정량과 근사적 계산(small-θ approximation)을 비교하기 위하여 one-sample case에서 몬테 카를로 모의실험을 통하여 변환 모수 λ가 0 근처의 수가 아니고, 표준편차 σ가 작을수록 추정량이 모수에 더 근사하고 위의 방법도 같아진다는 결론을 얻었다. 그밖에도 Manly(1976)는 관측치가 음수일 경우 변환을 이용하여 수정한 후 분석을 실시한 할 것을 제안하였고 Bickel과, Doksum(1981)은 support가 유한일 경우의 문제점을 또 다른 멱 변환을 사용하여 해결할 것을 제안하였다. 또한 Paulr과 Philip(1999)는 최적의 모형을 찾기 위한 변환이 최적의 예측을 얻기 위해 사용되어야만 하는 것은 아니라는 결론을 주었다. 시간적으로 서로 상관이 있는 자료에서도 분산이 일정하지 않을 경우에 이를 안정화 시켜 주는 방법으로 Box-Cox 멱 변환(1964)을 이용한다. 그러나 변환 모수(Transformation Parameter)를 정하는 것이 쉬운 일이 아니며 때에 따라서는 변환의 효과를 얻을 수 없는 경우가 발생한다. 신 기일과 강 희정(2001)은 평균에 비하여 분산이 상대적으로 작은 경우 ARMA모형의 차수, 모수 추정 및 예측 결과에 Box-Cox 변환이 영향을 미치지 않은 것을 밝혔다. 공간적으로 상관관계가 있는 자료를 분석함에 있어서도 긴 꼬리를 가진 비대칭형의 분포를 따를 경우 또는 이상치의 영향을 축소할 경우 자료의 변환은 많이 이용되고 있다.(Cressie, 1993) 본 논문에서는 변환이 공간자료에서 모수 추정에 어떤 영향을 미치는지를 알아보았다. 먼저 공간자료(Spatial data)를

1) 경기도 용인시 모현면 왕산리 한국외국어 대학교 통계학과 박사과정
E-mail : jhlee@stat.hufs.ac.kr
2) 경기도 용인시 모현면 왕산리 한국외국어 대학교 정보통계학과 부교수
E-mail : keyshin@stat.hufs.ac.kr

Z_X라하고 변환 모수 λ를 갖는 Box-Cox 변환을 고려한다고 하자. 그리고 변환된 자료를 $Y_\lambda(X)$라 하자.

$$Y_\lambda(X)=\begin{cases} Z^\lambda(X), & \lambda\neq 0 \\ \log(Z(X)), & \lambda=0 \end{cases} \tag{1}$$

또한 변환된 자료 $Y_\lambda(X)$는 대칭이고 분산이 일정하다고 가정하자. 그러면 우리는 $Y_\lambda(X)$ 자료를 이용하여 모수를 추정한 후 예측은 재변환(Retransformation)을 이용하게 된다.

본 논문에서는 이러한 결과가 공간자료분석에서 어떠한 영향을 보이는지 살펴보았다. 2절에서는 Box-Cox 변환이 Correlogram, Variogram에 어떠한 영향을 미치는지 살펴보았고 3절에서는 실제자료를 분석하여 2절에서 얻은 결론의 타당성을 뒷받침하였으며, 최종적인 결론은 4절에 있다.

2. 본론

서론에서 언급한 Box-Cox 변환을 고려하고, 신과 강(2001)에서처럼 Taylor 전개를 1차까지 고려하면

$$Y_\lambda(X) \approx \begin{cases} \mu_Z^\lambda+\lambda\mu_Z^{\lambda-1}(Z(X)-\mu_Z), & \lambda\neq 0 \\ \log(\mu_Z)+\mu_Z^{-1}(Z(X)-\mu_Z), & \lambda=0 \end{cases} \tag{2}$$

가 된다. 여기서 $\mu_Z=E(Z(X))$로 일정하다. 따라서 (2)의 결과를 이용하면

$$\rho_{Y_\lambda(X_1),Y_\lambda(X_2)} \approx \rho_{Z(X_1),Z(X_2)} \tag{3}$$

가 되어 Taylor 전개가 1차까지로 근사 될 경우 공간통계학의 Correlogram은 변환에 의해 영향을 받지 않게 된다.

다음으로 공간통계 모형설정 시 사용되는 Variogram을 살펴보자. 확률 부분을 모형화 하기 위하여 가장 많이 사용되는 방법은 먼저 자료에서 Variogram을 추정하고 추정된 결과를 이용하여 이론적으로 알려진 Variogram을 적합 시키는 것이다. 이때 일반적으로 사용되는 Variogram 추정량은 Matheron(1962)이 제안한 방법과 Cressie 와 Hawkins (1980) 방법 그리고 Genton (1998) 방법이 있다. 먼저 Matheron(1962) 방법을 살펴보면 다음과 같다.

$$2\hat{\gamma}_1(h)=\frac{1}{N(h)}\sum_{i=1}^{N(h)}(Y_\lambda(X_i)-Y_\lambda(X_i+h))^2 \tag{4}$$

만일 자료에 이상점이 존재할 경우 Materon의 classical한 방법보다는 Robust한 방법을 주로 사용하는데 대표적으로 사용되는 robust 추정량은 Cressie 와 Hawkins(1980)가 제안한 식(5)와 Genton(1998)이 제안한 추정량인 (6)식이다.

$$2\hat{\gamma}_2(h)=\frac{1}{0.457+0.494/N(h)}\left\{\frac{1}{N(h)}\sum_{i=1}^{N(h)}|Y_\lambda(X_i)-Y_\lambda(X_i+h)|^{1/2}\right\}^4 \tag{5}$$

$$2\hat{\gamma}_3(h)=[2.2191\{|V_i(h)-V_j(h)|;i<j\}_{(k)}]^2 \tag{6}$$

여기서 $k=\binom{[N(h)/2]+1}{2}$이고 $V_i=Y_\lambda(X_i)-Y_\lambda(X_i+h)$ 이다.

위의 (4)-(6)식 에서 Variogram은 $Y_\lambda(X_i)-Y_\lambda(X_i+h)$을 통하여 얻어지고 있으므로 이들을 변환하였을 경우의 영향력을 살펴보자.
(2)식을 이용하면 먼저 $\lambda\neq 0$인 경우

$$Y_\lambda(X_i)-Y_\lambda(X_i+h)\approx\lambda\mu_Z^{\lambda-1}\{(Z(X_i)-Z(X_i+h)\} \tag{7}$$

이 되고 $\lambda=0$인 경우는

$$Y_\lambda(X_i)-Y_\lambda(X_i+h)\approx\frac{1}{\mu_Z}\{(Z(X_i)-Z(X_i+h)\} \tag{8}$$

이 됨을 알 수 있다. 따라서 (7)식과 (8)식을 $\hat{\gamma}_i(h)$, $i=1,2,3$에 대입하게 되면 변환된 후의 Variogram은 다음과 같이 된다.

$$\hat{\gamma}_i^{Y}(h)\approx\begin{cases}(\lambda\mu_Z^{\lambda-1})^2\ \hat{\gamma}_i^{Z}(h), & \lambda\neq 0\\ \frac{1}{\mu_Z^2}\ \hat{\gamma}_i^{Z}(h), & \lambda=0\end{cases},\quad i=1,2,3$$

여기서 c를 다음과 같이 정의하면

$$c=\begin{cases}(\lambda\mu_Z^{\lambda-1})^2, & \lambda\neq 0\\ \frac{1}{\mu_Z^2}, & \lambda=0\end{cases} \tag{9}$$

위의 식은

$$\hat{\gamma}_i^{Y}(h)\approx c\cdot\ \hat{\gamma}_i^{Z}(h) \tag{10}$$

이 되어 변환된 Variogram의 추정값은 변환을 하지 않은 Variogram 추정값의 상수 배로 구해지게 된다. 다음에서 주로 사용되는 이론적인 Variogram의 예들인 Exponential Variogram, Gaussian Variogram 그리고 Spherical Variogram을 살펴보자. 먼저 이들의 Variogram들을 살펴보면 다음과 같다.

Spherical : $\gamma_0(t,\theta)=\begin{cases}0, & t=0\\ \theta_0+\theta_1\{\frac{3}{2}\frac{t}{R}-\frac{1}{2}(\frac{t}{R})^3\}, & 0<t\leq R\\ \theta_0+\theta_1, & t\geq R\end{cases}$

Exponential : $\gamma_0(t,\theta)=\begin{cases}0, & t=0\\ \theta_0+\theta_1(1-e^{-t/R}), & t>0\end{cases}$

Gaussian : $\gamma_0(t,\theta)=\begin{cases}0, & t=0\\ \theta_0+\theta_1(1-e^{-t^2/R^2}), & t>0\end{cases}$

여기서 θ_0는 Nugget을 θ_1은 Sill을 그리고 R은 Range를 나타낸다.

(10)에 의해서 우리는 쉽게 $\hat{\theta}_{0Y} \approx c \cdot \hat{\theta}_{0Z}$과 $\hat{\theta}_{1Y} \approx c \cdot \hat{\theta}_{1Z}$ 그리고 $\hat{R}_Y \approx \hat{R}_Z$의 추정량을 얻을 수 있다.

따라서 변환된 자료를 이용하여 추정된 Variogram은 변환 전 자료를 이용하여 추정한 자료의 상수 배로 구해지게 된다. 즉

$$\gamma_0(t, \hat{\theta}_Y) = c \cdot \gamma_0(t, \hat{\theta}_Z) \tag{11}$$

3. 자료 분석

다음 자료는 Jura 자료(A. G. Journel, 1997)로 이를 이용하여 본 논문의 이론에 대한 확인을 하였다.

λ가 1인 경우가 원 자료이고, 0.5인 경우가 제곱근 변환, 0인 경우가 log 변환이며, -0.5인 경우가 제곱근의 역 변환, -1.0인 경우가 역수 변환이다.

여러 모형 중 본 논문에서는 Spherical 모형을 이용하여 모수를 추정하였다. 본 논문의 가정이 평균이 크고 분산이 작은 경우이므로 원 자료에 각각 50. 100, 500 그리고 1000을 더하여 평균을 크게 한 후 원 자료에서의 추정결과와 변환 한 후의 추정 결과를 비교하였다. 다음에서 <표 1>이 range에 대한 자료분석 결과이고 <표2>가 sill에 대한 추정결과, <표3>이 nugget에 대한 추정결과이다. 또한 <그림 1>은 Spherical 모형을 이용하여 적합 시킨 Variogram 결과의 그림이고, <그림 2>가 Exponential 모형을 이용한 Variogram 적합 결과의 그림이다.

<표 1> 모수추정 결과(range)

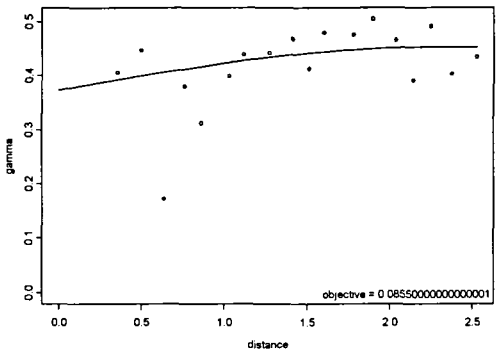
	원자료	원자료+50	원자료+100	원자료+500	원자료+500
1.0	2.2614	2.26184	2.26184	2.26183	2.26183
0.5	2.08392	2.25901	2.26066	2.26208	2.26211
0	1.97814	2.25564	2.25893	2.26178	2.26241
-0.5	1.842431	2.25244	2.25725	2.26144	2.2618
-1.0	1.67744	2.2446	2.25554	2.26098	2.26185

<표 2> 모수추정결과(sill)

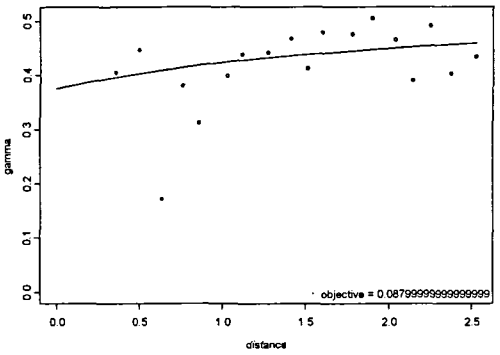
	원자료	원자료+50	원자료+100	원자료+500	원자료+500
1.0	0.0799216	0.079918009	0.079917448	0.0799178784	0.0799199206
0.5	0.01883249	0.000389748	0.000197297	3.98601e-005	1.99554e-005
0	0.0785196	3.04153e-005	7.79344e-006	3.18077e-007	7.96941e-008
-0.5	0.02280674	1.48354e-007	1.92413e-008	1.5864e-010	1.98998e-011
-1.0	0.11786	1.15837e-008	7.60088e-010	1.26584e-012	7.95042e-014

<표 3> 모수추정 결과(nugget)

	원자료	원자료+50	원자료+100	원자료+500	원자료+1000
1.0	0.372977	0.372985821	0.372986599	0.372984439	0.372986305
0.5	0.0796938	0.00181697	0.000920313	0.000186	9.31238e-005
0	0.292219	0.000141621	3.63325e-005	1.48412e-006	3.72021e-007
-0.5	0.0714769	6.89941e-007	8.96462e-008	7.40085e-010	9.28768e-011
-1.0	0.297817	5.37686e-008	3.53913e-009	5.90517e-012	3.71041e-013



<그림 1> spherical



<그림 2> exponential

분석결과를 살펴보면 본 자료에 어떠한 모형을 선택하였을 경우라도 1차 Taylor 전개가 가능할 경우 평균이 커질수록 변환 후의 range는 원 자료의 range에 수렴하고 sill과 nugget은 원 자료의 상수배 임을 확인 할 수 있다.

3. 결론

이상에서 Box-Cox 멱 변환이 공간통계 분석에 미치는 영향을 살펴보았다. 만약 분산이 평균에 비하여 상대적으로 작아서 Box-Cox 변환이 1차의 Taylor 전개로 근사 될 수 있을 경우 Correlogram은 변환에 의해 영향을 받지 않음을 알 수 있다. 또한 Variogram에 포함된 모수의 추정 값은 range의 경우 변환에 의해 영향을 받지 않고 sill과 nugget의 경우는 변환전과 변환 후 값의 상수 배 인 것으로 밝혀졌다. 자료분석에 있어 원 자료가 정상성을 만족하지 않을 때 특히 분산이 일정하지 않을 경우 우리는 변환을 하게 되는데, 이런 변환은 분산을 안정화시킴으로서 추정 및 검정에 타당성을 주는 반면 새로운 편의를 생성하거나(Granger & Newbold,1976) 모형을 복잡하게 만듦으로써 해석의 어려움도 수반한다. 그러므로 본 논문에서의 결론에 따라 공간통계분석에서의 모형 설정에 있어 본 논문의 가정에 맞는 자료라면 위 결과를 이용할 수 있을 것이다.

참고문헌

A. G. Journel (1997), Geostatistics for natural resources evaluation, Oxford University Press.

Bickel, P. J. and Doksum, K. A. (1981), An analysis of Transformations Revisited. Journal of the American Statistical Association, Vol. 76. 296-311.

Box, G. E. P. and Cox, D. R. (1964), An analysis of transformations, Journal of the Royal Statistical Society, Series B, Vol. 26, 211-252.

Cressie, N. and Hawkins D. M. (1980), Robust estimation of the variogram, I. Mathematical Geology, Vol. 12, No. 2, 115-125.

Cressie, N. (1993), Statistics for spatial data, John Wiley & Sons, Inc.

Genton, M. C. (1998), Highly robust variogram estimation, Mathematical Geology, Vol. 30, No. 2, 213-221.

Jeremy, M. G. Taylor (1986), The retransformed mean after a fitted power transformation. Journal of the American Statistical association, Vol. 81, No. 393, pp 114-118.

Paulr, D. B. , Philip, H. F. (1999), Forecasting power-transformed time series data. Journal of Applied Statistics, Vol. 26. No. 7, pp 807-815.

Granger, C. W. J. and Newbold, P. (1976), Forecasting Transformed Series. Journal of the Royal Statistical Society, Series B(Methodological), Vol. 38, Issue 2, 189-203.

Mainly, B. F. (1992), Exponential daata transformation. The Statistician, Vol. 25, 37-42.

Matheron, G. (1962), Traite de Geostatistique appliquee, Tome I. Memoires du Bureau de Recherches Geologiques et Minieres, No. 14. Editions Technip, Paris.

Shin, K-I., and H-J. Kang (2001), A study on the effect of power transformation in the ARMA(p,q) model, Journal of Applied Statistics, Vol. 28, No. 8, 1019-1028.

Stephen, P. Kaluzny, Silvia C. Vega, Tamre P. Cardoso and Alice, A. Shelly (1998), S+SpatialStats User's Manual, Springer.