응용통계연구 제 12권 2호, 1999년, pp. 347-362

고차의 마코브 연쇄-의존 모델을 이용한 남한 강수량 자료의 분석*

박정수[1] 정영근[2] 김래선[3]

요 약

강수 형태 및 강수량을 동시에 고려하는 1차의 마코브 연쇄-의존 모델을 고차의 모델로 확장하였다. 남한의 53개 지역의 강수량 자료에 대해 계절별로 마코브 연쇄의 차수를 결정하였고, 고차의 마코브 연쇄-의존 모델을 적용하여 강수량의 분포 특성을 살펴 보았다.

1. 동기 및 기본 개념

마코브 연쇄의 대표적인 예가 강수 형태일 것이다. 실제로 Moon, Ryoo and Kwon(1994)은 남한의 강수 자료에 대해 마코브 연쇄를 적용하여 강수 형태 등을 파악하였다. 그림 1.1은 1990년 6월부터 8월까지 전주지역에 대한 강수그림이다. 막대의 크기는 강수량을 나타낸다. 그림으로 보아 강수형태는 마코프 연쇄를 따를 것으로 예상된다. 그런데 마코브 연쇄는 비가 왔는가 안 왔는가에 관심이 있지 비온 날의 강수량은 고려하지 못한다. 따라서 강수형태와 강수량을 동시에 고려하는 모델링을 해보자는 것이 본 연구의 출발이였다. 즉, 강수량은 날자에 따라 독립적으로 분포하는 것이 아니라 강수 형태에 따라 영향을 받는다는 것이다. 이에 대해 Katz(1977)는 O'Brien(1974) 등에 의해 연구된 연쇄-의존 모델의 점근 분포이론을 적용하여 강수량 분석을 하였다. 그런데 Katz(1977) 는 1차 마코프 연쇄를 이용하였으나, 우리가 한국 강수 자료에 대해 마코프 연쇄의 차수를 추정해 본 바 특히 봄철에 2차의 마코브 연쇄를 보이는 곳이 많았다. 따라서 본 논문에서는 1차의 마코브 연쇄-의존 모델을 고차의 모델로 확장하고, 이를 남한의 강수량 자료에 적용한다.

1.1. 마코브 연쇄

가산적 상태공간 $S = \{0, 1, 2, \cdots, s-1\}$를 갖는 확률과정 $\{X_t,\ t = 0, 1, 2, \cdots\}$ 을 생각하자. 예를 들어 강수 형태와 관련하여 X_t 를 다음과 같이 정의할 수 있다.

$$X_t = \begin{cases} 0, & t \text{ 번째 날에 맑거나 흐린 경우,} \\ 1, & t \text{ 번째 날에 비온 경우.} \end{cases}$$

* 이 논문은 1998년도 전남대학교 학술연구비 지원에 의하여 연구되었음.

1) (500-757) 광주광역시 북구 용봉동 300, 전남대학교 기초과학연구소 및 자연과학대학 통계학과, 교수
2) (500-757) 광주광역시 북구 용봉동 300, 전남대학교 사범대학 지구과학교육과, 교수
3) (500-757) 광주광역시 북구 용봉동 300, 전남대학교 자연과학대학 통계학과, 대학원 졸업

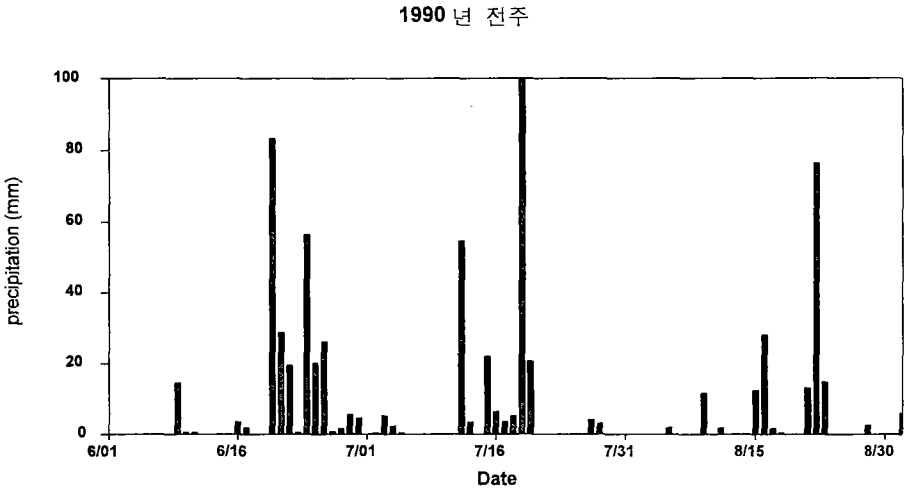

그림 1.1: 1990년 여름의 전주지역에 대한 강수량의 막대그래프

1차 마코브 연쇄(first-order Markov chain)의 경우에 X_t는 다음과 같은 가정을 만족하는 확률과정으로 정의된다.

$$P(X_t = j|X_{t-1} = i, X_{t-2}, \cdots, X_0) = P(X_t = j|X_{t-1} = i) = p_{ij}.$$

이들 1차 추이확률 p_{ij} 들로 이루어진 행렬 P를 추이 행렬이라고 한다. 본 논문에서는 또한 정상 (stationary) 마코브 연쇄를 가정한다.

자료가 주어지면 이들 추이확률은 최우추정법에 의해서

$$\hat{p}_{ij} = \frac{n_{ij}}{n_i} \quad i, j = 0, 1, 2, \cdots, s-1$$

으로 추정된다. 여기서 n_i 는 상태 i 를 갖는 관측치의 수이고, n_{ij} 는 현 시점에서 상태 i 를 가지고 다음 시점에서 상태 j 로 변화한 횟수이다. 마코브 연쇄에 대한 자세한 내용은 Ross(1985) 나 Guttorp(1995) 또는 Kemeny and Snell(1983) 을 보기 바란다.

2차(second order) 마코프 연쇄 $\{X_t\}$는 다음과 같은 가정으로 정의된다.

$$\begin{aligned} &P[X_t = j|X_{t-1} = i, X_{t-2} = h, X_{t-3}, \cdots] \\ &\quad = P[X_t = j|X_{t-1} = i, X_{t-2} = h] \ = p_{hij}. \end{aligned}$$

나아가 r차 마코프 연쇄 $\{X_t\}$ 는 다음과 같은 가정으로 정의된다.

$$\begin{aligned} &P(X_{t+1} = i_{r+1}|X_t = i_r, X_{t-1} = i_{r-1}, \cdots, X_0) \\ &\quad = P(X_{t+1} = i_{r+1}|X_t = i_r, X_{t-1} = i_{r-1}, \cdots, X_{t-r+1} = i_1) \\ &\quad = p_{i_1, i_2, \cdots, i_{r+1}}. \end{aligned}$$

마코프 연쇄의 차수가 높아질수록 추이확률은 복잡한 형태가 된다.

1.2. 마코브 연쇄의 차수 결정

마코프 연쇄의 차수를 결정하는 방법으로 보통 AIC 또는 BIC 가 이용된다. 먼저 AIC는 k차수의 마코프 연쇄에 대한 아래와 같은 최대우도 함수의 계산이 요구된다(상수 항은 제외).

$$M_k(X_1, \cdots, X_n) = \prod_{i_1, \cdots, i_{k+1}} (\hat{p}_{i_1, \cdots, i_{k+1}})^{n_{i_1, \cdots, i_{k+1}}}.$$

여기서,

$$\hat{p}_{i_1, \cdots, i_{k+1}} = \frac{n_{i_1, \cdots, i_{k+1}}}{n_{i_1, \cdots, i_k}}$$

이며, $n_{i_1, \cdots, i_{k+1}}$는 $i_1 \rightarrow i_2 \rightarrow \cdots \rightarrow i_{k+1}$ 이 되는 변화의 수이다. 마코프 연쇄 차수의 가상적인 상한 m에 대하여 k차수를 검정하려 할 때, 우도비 검정통계량에 근거한 AIC(Akaike's Information Criterion)는, 상태의 개수 s 에 대하여

$$\mathrm{AIC}(k) = 2 \ln \lambda_{k,m} + 2(s^m - s^k)(s-1).$$

이다 (Katz, 1981). 여기서

$$\lambda_{k,m} = \frac{M_k(X_1, \cdots, X_n)}{M_m(X_1, \cdots, X_n)}$$

이다. 우리는 AIC(k)를 최대로 하는 k를 찾는다.

다음으로 BIC(Bayes Information Criterion) 는

$$\mathrm{BIC}(k) = 2 \ln \lambda_{k,m} + (s^m - s^k)(s-1)\, log\, n$$

인데, 우리는 BIC(k) 을 최대로 하는 k를 찾는다. 본 논문에서는 0차(독립 표본이라는 가정)에서 3차까지 고려하여 차수를 정했다 (즉, m=3).

실제 계산에 있어서는, 3차는 n개의 관찰값 중에서 마지막 $n-3$개의 관찰값만을 사용한다. 그리고 2차는 $n-2$개의 관찰값만을, 그리고 1차는 $n-1$개의 관찰값을 사용한다. 왜냐하면, 3차일 경우 4번째 관찰값부터 계산을 하기 때문이다. 처음 3개의 관찰값은 변화하는 상태를 관찰하는데만 쓰인다.

BIC는 Guttorp(1995) 등에 의해서 추천되고 있으며, 마코프 연쇄의 차수에 대해 일치 추정량이므로(Katz, 1981) 본 논문에서도 BIC를 이용하여 차수를 결정한다. 실제로 BIC(k)를 계산할 때 우리는 위의 정의와 같은 의미를 갖는

$$\mathrm{BIC}(k) = 2\, l(\hat{\theta}) - k\, log\, n$$

를 이용하였다 (Guttorp, 1995). 여기서 $l(\hat{\theta})$ 는 최대 로그우도 함수로서 차수가 3인 경우 $l(\hat{\theta}) = \sum\sum\sum n_{hij}\, log\, \hat{p}_{hij}$ 이며, n은 각 차수마다 일정한 값으로 n_{hij} 의 합이다.

2. 고차의 마코브 연쇄를 1차의 형태로 표현

고차의 마코브 연쇄를 일차의 마코브 연쇄로 표현할 수 있음을 기술하겠다. 여기서는 0(비 안옴)과 1(비 옴) 두가지 상태를 갖는 경우만 생각한다.

만약 $\{X_t\}$ 가 2차라면 새로운 과정 $\{Y_t\}$ 를 다음과 같이 정의한다.

$$Y_1 = (X_0, X_1), \quad Y_2 = (X_1, X_2), \quad \cdots, \quad Y_n = (X_{n-1}, X_n).$$

그리고 Y_t 가 가질 수 있는 값이 $(0,0)$, $(0,1)$, $(1,0)$, $(1,1)$ 인데 이들을 각각 0, 1, 2, 3 으로 변환시켜 준다. 그러면 $\{Y_t\}$ 는 상태가 4가지인 1차의 마코브 연쇄가 된다. 그런데 특기할 점은 $\{Y_t\}$ 에 대한 추이확률은 Y_n 과 Y_{n+1} 의 정의로부터 Y_n 에서 X_n 과 Y_{n+1} 에서 X_n 이 같은 값을 가질 때만 존재하게 된다. 그 외의 경우에는 추이확률이 모두 0 이다.

만약 $\{X_t\}$ 가 3차라면 새로운 과정 $\{Y_t\}$ 를

$$Y_t = (X_{t-2}, X_{t-1}, X_t), \quad Y_{t+1} = (X_{t-1}, X_t, X_{t+1}), \quad \cdots.$$

로 정의하고 상태값을 적절히 변환해 줌으로서 (7개의 상태 값을 갖는) 1차 마코브 연쇄로 표현할 수 있다. 일반적으로 r차인 경우에 다음과 같이 표현된다.

$$Y_t = (X_{t-r+1}, \cdots, X_{t-1}, X_t), \quad Y_{t+1} = (X_{t-r+2}, \cdots, X_t, X_{t+1}), \quad \cdots.$$

상태값의 변환은 2진법을 사용하여 간단히 이루어진다. 예를 들어 2차인 경우, $(X_0 = 1, X_1 = 0)$ 이면 $\{Y_1 = (X_0 = 1, X_1 = 0)\} = 0 \cdot 2^0 + 1 \cdot 2^1 = 2$의 값을 갖는다. 만약 $(X_0 = 1, X_1 = 1)$ 이면 $\{Y_1 = (X_0 = 1, X_1 = 1)\} = 0 \cdot 2^1 + 1 \cdot 2^1 = 3$의 값을 갖는다. 그리고 3차인 경우, $(X_0 = 1, X_1 = 0, X_2 = 1)$ 이면 $\{Y_2 = (X_0 = 1, X_1 = 0, X_2 = 1)\} = 1 \cdot 2^0 + 0 \cdot 2^1 + 1 \cdot 2^2 = 5$의 값을 갖게 된다. 따라서 2차인 경우 Y_n이 가질 수 있는 값은 $\{0, 1, 2, 3\}$ 이고, 3차인 경우는 $\{0, 1, 2, \cdots, 7\}$이다. 나아가 r차인 경우에는 $\{0, 1, 2, \cdots, 2^r - 1\}$이다.

일반적으로 표현하면, X_t 가 갖는 값을 d_t, X_{t-1} 이 갖는 값을 d_{t-1}이라 하고, X_{t-r+1} 이 갖는 값을 d_{t-r+1} 이라고 할 때,

$$Y_n \text{ 의 상태 } = d_n 2^0 + d_{n-1} 2^1 + \cdots + d_{n-r+1} 2^{r-1}$$

으로 변환하여 표기할 수 있다. 이와 같이 고차의 마코브 연쇄를 일차의 마코브 연쇄 형태로의 표현은 다음 절에서 다루는 연쇄-종속 모델을 고차의 경우로 확장하는데 매우 중요하게 이용된다.

3. 마코브 연쇄-의존 모델

3.1. 1차 마코브 연쇄 경우

강수 형태와 강수량을 동시에 고려하기 위해 먼저, Z_t 를 t 날의 강수량이며, Z_t 는 $X_t = 1$ 일 때 0보다 큰 값을 가지며 $X_t = 0$ 일 때는 0을 갖는다. 또 $\{X_t\}$는 추이확률이 $\{p_{ij}\}$, 정상

확률 π를 갖는 1차 마코프 연쇄라고 하자. 우리의 관심은 n 일 간의 강수량이며 이것은

$$S_n = \sum_{t=1}^{n} Z_t$$

로 정의된다. 우리가 알려고 하는 것은 S_n 의 확률분포 즉 $F_{S_n}(s) = P(S_n < s)$ 이다. 여기서 Z_t 는 독립적인 변수가 아니고 마코브 연쇄인 $\{X_t\}$ 에 의존하기 때문에 S_n 를 마코브 연쇄-의존 과정(chain-dependent process)이라고 부른다. 이 연쇄-의존 모델은 다음 두 개의 가정을 한다.

가정1. Z_t 는 (X_t, X_{t-1})에 의존한다.
가정2. $\{X_t\}_{t=0,1,\cdots}$가 주어진 조건에서 $\{Z_t\}_{t=1,2,\cdots}$ 는 독립이다.

위의 가정으로부터 우리는 다음을 얻는다.

$$\begin{aligned} &P(Z_t \le z | X_0, Z_1, X_1, Z_2, X_2, \cdots, Z_{t-1}, X_{t-1} = i) \\ &\quad = P(Z_t \le z | X_{t-1} = i). \end{aligned}$$

또한 오늘 비가 왔을 때 어제의 상태를 고려한 오늘 강수량의 조건부 평균과 조건부 분산을 $i = 0,\ 1$ 에 대해 다음과 같이 표기하자.

$$\mu_{i1} = E(Z_t | X_{t-1} = i, X_t = 1), \tag{3.1}$$

$$\sigma_{i1}^2 = Var(Z_t | X_{t-1} = i, X_t = 1). \tag{3.2}$$

이제 n 날 간의 강수량 S_n 에 대해서 O'Brien(1974)과 Katz(1977)는 다음 극한정리를 얻었다.

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{d} N(0,1). \tag{3.3}$$

식 (3.3)에서 μ 와 σ^2 는

$$\mu = \sum_i \pi_i \, \mu_{i1} \, p_{i1},$$

$$\sigma^2 = \rho_0 + 2\pi Q \left([I - (P - \Pi)]^{-1} - \Pi\right) Qe$$

이다. 여기서 $\rho_0 = Var(Z_t)$, $\pi = (\pi_0 \ \pi_1)$, $\Pi = e\pi$, $e^t = (1\ 1)$, $Q = \{p_{ij} \ \mu_{ij}\}$ 이다. 구체적 유도 과정은 부록에 자세히 기술되었다.

3.2. 고차의 연쇄-의존 모델

이제 고차의 마코브 연쇄에 대한 연쇄-의존 모델을 고려하자. $\{X_t\}$ 가 r차의 마코브 연쇄라고 할 때 제2장에서 기술된 바와 같이 새로운 과정 $\{Y_t\}$ 를 정의하면 이것은 2^r 개의 상태를 갖는 1차의 마코브 연쇄가 된다. 따라서 S_n의 확률분포는 앞에서의 결과에 $(Z_t,\ X_t)$

대신에 $(Z_t,\ Y_t)$를 고려함으로서 구해진다. 다만 조건부 평균 및 분산은 다음과 같이 약간 다르게 구해진다.

$$\begin{aligned}\mu_{ij} &= E[Z_t|(X_{t-r},\cdots,X_{t-2},X_{t-1}),X_t=1]\\ &= E(Z_t|Y_{t-1}=i,Y_t=j), \qquad j \text{ 는 홀수}\\ \sigma_{ij}^2 &= Var(Z_t|Y_{t-1}=i,Y_t=j)), \quad j \text{ 는 홀수}.\end{aligned}$$

여기서 j 가 0 또는 짝수인 경우 $\mu_{ij}=0,\ \sigma_{ij}^2=0$ 이다. 그 이유는, μ_{ij} 와 σ_{ij}^2 은 $X_t=1$ 일 때만 정의되는데, 이때 변환된 Y_t 의 상태의 값이 홀수를 갖기 때문이다.

이제 1차 마코브 연쇄-의존 모델에서와 마찬가지로 고차 모델의 경우에도 식 (3.3)과 같은 극한정리를 얻게 되는데, 여기서의 μ와 σ를 부록의 유도과정(1차의 경우)을 이용하여 표현하면 다음과 같다.

$$\begin{aligned}\mu &= E(Z_t) = E[\ E(Z_t|Y_{t-1}=i,Y_t=j)\ ]\\ &= \sum_{i=0}^{2^r-1}\sum_{j=0}^{2^r-1}\pi_i\ \mu_{ij}\ p_{ij} \ = \ \pi Q e,\\ \sigma^2 &= \rho_0 + 2\sum_{j=1}^{\infty}\rho_j\\ &= \rho_0 + 2\pi Q\ ([I-(P-\Pi)]^{-1}-\Pi)\ Qe,\\ \rho_0 &= Var(Z_t)\\ &= \sum_{i\,j}\pi_i\ \sigma_{ij}^2\ p_{ij} + \sum_{i\,j}\pi_i\ \mu_{ij}^2\ p_{ij} - \mu^2\\ &= \pi Q_{(\sigma^2)}e + \pi Q_{(\mu^2)}e - \mu^2.\end{aligned}$$

위에서 쓰인 행렬들은 Y_t 의 상태의 수 $m=2^r$ 에 대해서 다음과 같다.

$$I: m\times m \text{ 단위 행렬},\ Q=\{p_{ij}\ \mu_{ij}\},\ Q_{(\sigma^2)}=\{p_{ij}\ \sigma_{ij}^2\},$$

$$Q_{(\mu^2)}=\{p_{ij}\ \mu_{ij}^2\},\ \Pi=e\pi,\ e=\mathbf{1}_m \text{ (단위 벡터)}.$$

3.3. 2차 마코브 연쇄-의존 모델의 경우

예를 들어 2차 마코브 연쇄의 경우 식 (3.3) 극한정리의 μ와 σ은 다음과 같다. j 가 0 또는 짝수인 경우 $\mu_{ij}=0,\ \sigma_{ij}^2=0$ 이다는 점을 이용하여,

$$\begin{aligned}\mu &= \sum_{i=0}^{3}\sum_{j=0}^{3}\pi_i\ \mu_{ij}\ p_{ij}\\ &= \pi_0\mu_{01}p_{01}+\pi_1\mu_{13}p_{13}+\pi_2\mu_{21}p_{21}+\pi_3\mu_{33}p_{33},\\ \sigma^2 &= \rho_0 + 2\pi Q\ ([I-(P-\Pi)]^{-1}-\Pi)\ Qe.\end{aligned}$$

여기서 $\Pi = e\pi$ 이고,

$$\begin{aligned}\rho_0 &= \sum_{i\,j} \pi_i\, \sigma_{ij}^2\, p_{ij} + \sum_{i\,j} \pi_i\, \mu_{ij}^2\, p_{ij} - \mu^2 \\ &= \pi_0 p_{01}(\sigma_{01}^2 + \mu_{01}^2) + \pi_1 p_{13}(\sigma_{13}^2 + \mu_{13}^2) \\ &\quad + \pi_2 p_{21}(\sigma_{21}^2 + \mu_{21}^2) + \pi_3 p_{33}(\sigma_{33}^2 + \mu_{33}^2) - \mu^2.\end{aligned}$$

이다. 위에 쓰인 행렬 P와 (μ_{ij}) 및 (σ_{ij}^2)의 구체적인 모습은 다음과 같다.

$$P = (p_{ij}) = \begin{bmatrix} p_{00} & p_{01} & 0 & 0 \\ 0 & 0 & p_{12} & p_{13} \\ p_{20} & p_{21} & 0 & 0 \\ 0 & 0 & p_{32} & p_{33} \end{bmatrix},$$

$$(\mu_{ij}) = \begin{bmatrix} 0 & \mu_{01} & 0 & 0 \\ 0 & 0 & 0 & \mu_{13} \\ 0 & \mu_{21} & 0 & 0 \\ 0 & 0 & 0 & \mu_{33} \end{bmatrix}, \quad (\sigma_{ij}^2) = \begin{bmatrix} 0 & \sigma_{01}^2 & 0 & 0 \\ 0 & 0 & 0 & \sigma_{13}^2 \\ 0 & \sigma_{21}^2 & 0 & 0 \\ 0 & 0 & 0 & \sigma_{33}^2 \end{bmatrix}.$$

이들은 모두 자료로부터 추정된다. 또한 정상확률 $\pi = [\pi_0\ \pi_1\ \pi_2\ \pi_3]$ 는 $\pi P = \pi$, $\sum_i \pi_i = 1$ 의 해로서 구해진다.

4. 남한 강수량 자료에의 적용

4.1. 마코브 연쇄의 차수 결정

남한의 53개 지역에서 1980년부터 1994년까지 15년간의 매일의 강수량 자료를 분석에 사용하였다. 하루의 측정한 강수량이 $0.1mm$ 이상이면 비 온 것으로 하였다. 이들 자료를 계절별로 나눠서 BIC를 이용하여 마코브 연쇄의 차수를 결정하였다. 표 4.1은 그 중 몇 개의 지역에 대한 계절별 BIC와 선택된 차수이다.

전체 53개 지역에 대한 계절별 차수를 정리하면 다음과 같다.

봄: 지역에 따라 차이 (1차 또는 2차).

여름: 대부분 1차 (예외: 강화, 서산, 서귀포 → 2차).

가을: 전국이 1차.

겨울: 대부분 1차

(예외: 합천 → 2차, 울진, 부산, 강릉 → 미미한 차이로 2차).

여름, 가을, 겨울은 대부분 1차가 선택되었으나, 봄에는 지역에 따라 차이를 보였다. 그림 4.1은 봄에 선택된 차수를 지도에 표시한 것이다. 이 지도에서 알 수 있듯이, 서울, 경기, 강원 지역은 대부분 1차, 충청, 경상남북도, 전라남도, 제주도는 2차, 전라북도는 1차의 마코프 연쇄가 선택되었다.

표 4.1: 남한 일부 지역 강수량 자료에 대한 계절별 마코브 연쇄 차수 결정을 위한 BIC 값과 선택된 차수

지역	계절	BIC(0)	BIC(1)	BIC(2)	BIC(3)	선택
서울	봄	-1568.00	**-1476.70**	-1477.81	-1506.43	1
	여름	-1893.52	**-1704.50**	-1713.30	-1727.76	1
	가을	-1561.32	**-1454.29**	-1465.40	-1486.66	1
	겨울	-1479.32	**-1425.23**	-1433.09	-1445.01	1
광주	봄	-1701.60	-1586.22	**-1579.09**	-1605.91	2
	여름	-1889.45	**-1714.05**	-1725.66	-1750.31	1
	가을	-1629.37	**-1480.21**	-1490.47	-1516.99	1
	겨울	-1767.67	**-1653.50**	-1667.23	-1693.31	1
강릉	봄	-1643.03	-1513.52	**-1510.17**	-1537.001	2
	여름	-1906.49	**-1708.63**	-1716.10	-1737.20	1
	가을	-1582.23	**-1394.49**	-1407.23	-1427.87	1
	겨울	-1289.43	-1141.53	**-1141.25**	-1168.76	2(1)
강화	봄	-1465.03	**-1389.04**	-1393.15	-1420.50	1
	여름	-1835.68	-1681.40	**-1669.98**	-1689.26	2
	가을	-1457.48	**-1343.69**	-1356.19	-1381.74	1
	겨울	-1355.96	**-1301.40**	-1314.94	-1336.96	1
부산	봄	-1670.60	-1596.00	**-1585.98**	-1604.46	2
	여름	-1826.37	**-1618.59**	-1623.01	-1645.33	1
	가을	-1416.52	**-1280.96**	-1293.97	-1320.04	1
	겨울	-1280.37	-1171.82	**-1170.75**	-1194.17	2(1)

4.2. 마코브 연쇄의 차수 분포에 대한 기상학적 해석

마코브 연쇄의 차수의 계절별 차이에 대해, 특히 봄철에 2차 마코프 연쇄가 많이 나타난 것에 대해 기상학적인 해석을 하면 다음과 같다. 한반도의 일기는 중위도 상공에 형성되는 편서풍의 영향을 받아 6-7일 주기로 반복되는 경향이 있다. 특히 봄철은 상층 편서풍이 약화되는 시기로 편서풍 파동의 진폭이 줄어들면서 일기 변화의 규칙성이 커진다. 반면, 가을과 겨울은 시베리아기단이 급격히 형성 또는 발달되면서 간헐적으로 그 세력이 동아시아 지역으로 크게 확장되므로 일기의 변동폭이 크고 불규칙한 편이다. 여름은 대기 불안정에 의한 단속적인 기단성 강우가 많다. 그러므로 강수 발생에 대한 마코브 연쇄의 계절별 차이는 이러한 일기 변화의 규칙성을 반영하고 있는 것으로 해석된다.

봄철은 한반도 남쪽으로부터 온난하고 습윤한 기단 세력이 북쪽으로 확장되고 상승 제트기류도 약화되면서 북상하므로 기압골은 가을 및 겨울에 비해 동서방향 성분을 더 갖게 된다. 그러므로 봄철 강수는 지속성이 비교적 크고 규칙적인 반복성을 갖는 경향이 있다.

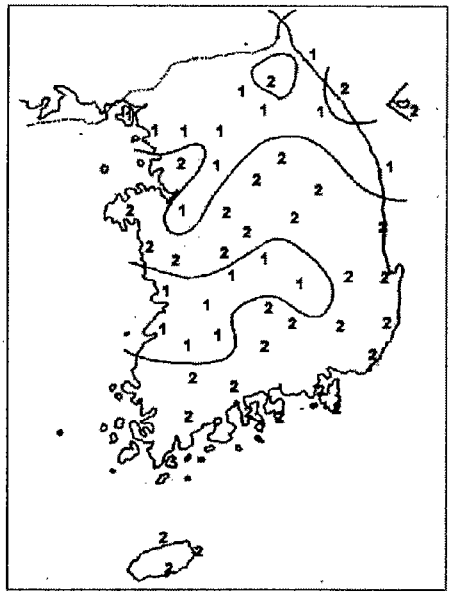

그림 4.1: 남한의 봄에 선택된 마코프 연쇄의 차수 분포 지도 (차수는 BIC 에 의해서 결정됨.)

봄철 남한의 대부분의 지역에서 강수 발생은 마코브 연쇄에서 2차의 차수를 나타내며 그 분포(그림 4.1)는 편서풍 파동에 연관하여 한반도를 지나는 저기압의 통과 경로 즉, 화북으로부터 한반도 중부 및 동지나해로부터 남부를 향하는 두 경로를 반영하고 있는 것으로 해석된다 (추춘기, 민경덕, 1983).

4.3. 차수 1차와 2차의 차이에 대한 확률적 해석

1차일 때 상태 1에서 다음 단계에 상태 0으로 갈 추이확률 p_{10} 는 2차일 때의 두 개의 추이확률 p_{010} 와 p_{110} 의 선형결합, $p_{10} = \pi_0\, p_{010} + \pi_1\, p_{110}$, 으로 표현된다. 이러한 표현으로부터 우리는 차수 선택에 대한 해석을 가할 수 있다. 즉, p_{000} 와 p_{100} 의 차이가 상대적으로 크면 2차 모형이 적합하고 그 차이가 작으면 1차 모형이 적합하다고 해석된다. 예를 들어 광주의 봄에는 2차가, 여름에는 1차가 선택되었는데 봄의 추이확률을 보면 다음과 같다.

1. 1차로 적합할 경우
   $p_{00} = .7644$ $\quad$ $p_{01} = .2355$
   $p_{10} = \mathbf{.5486}$ $\quad$ $p_{11} = .4513$

2. 2차로 적합할 경우
   $p_{000} = .7493$ $\quad$ $p_{001} = .2507$
   $p_{010} = \mathbf{.4545}$ $\quad$ $p_{011} = .5454$

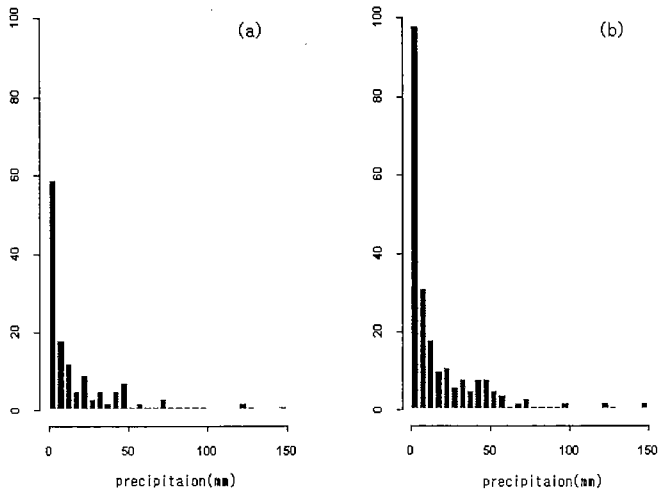

그림 4.2: 강화 지역의 봄: (a)전날 비가 안 왔을 경우 당일의 강수량과 (b)전날 비가 왔을 경우 당일의 강수량.

$p_{100} = .8125$ $p_{101} = .1875$
$p_{110} = \mathbf{.6629}$ $p_{111} = .3370$

여기서 예를 들어 p_{010} 와 p_{110} 는 .2084의 차이를 보인다. 이 차이가 상대적으로 크므로, p_{10} 보다는 p_{010} 와 p_{110} 으로 나누어 설명하는 것이 더 정확한 해석이 될 것이다. 따라서 봄에는 2차가 선택된 것이다. 한편 광주의 여름의 경우 예를 들어 p_{010} 와 p_{110} 의 차이는 .0281 인데 (실제값 생략) 이 차이는 상대적으로 작기 때문에 p_{10} 으로도 충분히 설명이 가능하다. 따라서 여름에는 1차가 선택된 것이다.

표 4.1: 서산 지역 봄철 (2차 연쇄-의존 모델)의 평균 강수량 및 표준편차

상태	평균	표준편차
$(X_{n-2}=0, X_{n-1}=0, X_n=1)$	9.02	12.40
$(X_{n-2}=0, X_{n-1}=1, X_n=1)$	11.57	15.76
$(X_{n-2}=1, X_{n-1}=0, X_n=1)$	5.30	7.81
$(X_{n-2}=1, X_{n-1}=1, X_n=1)$	10.72	21.65

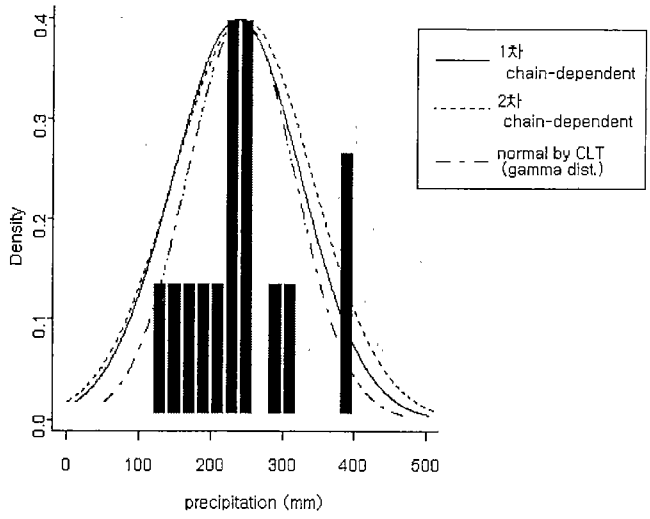

그림 4.3: BIC에 의해서 2차 마코브 연쇄-의존 모델이 선택된 봄철 광주의 92일간(3개월)의 강수량의 분포 (표본의 크기=15)

4.4. 연쇄-의존 모델의 적용

이제 강수량의 합에 대해 살펴보자. 먼저 몇 지역에 대해 강수량의 조건부 평균(μ_{ij}) 및 표준편차(σ_{ij})를 알아보면 다음과 같다. 표 4.2 은 서산 지역 봄철에 전날 및 전전날의 비온 상태에 따른 당일의 평균 강수량 및 표준편차이다.

여기서 $(X_{n-2}=1, X_{n-1}=1, X_n=1)$의 상태의 평균 강수량이 $(X_{n-2}=1, X_{n-1}=0, X_n=1)$의 상태보다 더 많고 표준편차도 크다는 것을 알 수 있다. 다시 말해서 삼일 연속 비가 왔을 경우에 삼일째의 강수량은 첫째날 비가 오고 둘째날 비가 안 오고 삼일째에 비온 경우의 삼일째의 강수량보다 더 많다는 것이다.

그림 4.2은 강화 지역의 봄(1차 연쇄-의존 모델)에 $(X_{n-1}=0, X_n=1)$의 경우와 $(X_{n-1}=1, X_n=1)$의 경우 강수량의 분포이다. 그림에서 볼 수 있듯이 봄철에 강화에서는 $(X_{n-1}=0, X_n=1)$ 보다 $(X_{n-1}=1, X_n=1)$ 인 경우에 비가 오는 횟수도 많고 양도 많다는 것을 알 수 있다.

4.5. 강수량의 점근 정규분포

이제 연쇄-종속 모델에서 S_n 의 점근 정규분포를 적용해 보자. 광주 지역의 봄에는 2차 마코프 연쇄가 선택되었으므로 3개월간(즉 3월, 4월, 5월의 92일간)의 강수량에 대해 적용해 보았다. 우리는 15년의 자료를 이용하였으므로 여기서 표본의 크기는 15이다. 자료로부터 추정된 추이확률을 바탕으로 앞 장에서 제시한 행렬 및 통계량들을 계산하여, 결과적으

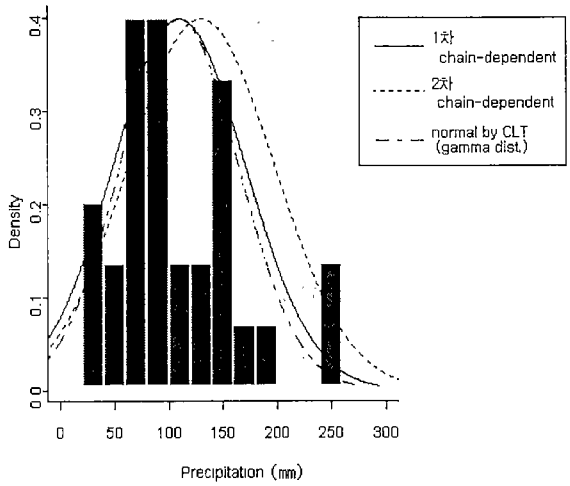

그림 4.4: BIC에 의해서 2차 마코브 연쇄-의존 모델이 선택된 봄철 서산의 46일간(1.5개월)의 강수량의 분포 (표본의 크기=30)

로 $\mu = 2.7093$, $\sigma = 10.2327$ 얻었다. 따라서 $S_n = \sum_1^{92} Z_i$ 는 식 (3.3)에 의해서 근사적으로 $N(249.26,\ (98.15)^2)$ 분포를 갖게 된다.

또한 비교를 위하여 독립 변수 및 1차 마코브 연쇄-의존 모델에 대해서도 같은 값들을 구해 보았다. 먼저 1차 연쇄-의존 모델의 경우 $\mu = 2.620$, $\sigma = 9.4183$ 이여서, $S_n \sim AN(241.04,\ (90.34)^2)$ 이다. 한편 Z_t 가 독립 변수일 때는 일반적으로 감마분포를 따른다고 알려져 있으므로(Suzuki, 1980) S_n 은 감마변수의 합이 되어 역시 감마분포를 따르게 된다. 감마분포의 모수를 자료로부터 추정하여 얻은 $\hat{\mu}$ 와 $\hat{\sigma}$ 는 각각 $\bar{X} = 241.087$, $S = 75.636$ 이다. 여기서는 비교를 위하여 대표본 경우의 정규분포로 근사시켰다.

그림 4.3은 광주 지역의 봄의 92일간의 강수량의 합에 대한 히스토그램과 1차 연쇄-의존 모델의 적용해서 구한 근사적 정규분포, 2차를 적용해서 구한 근사적 정규분포, 그리고 감마 분포의 합에 의한 근사적 정규분포 그림이다. 세가지 모형에서 큰 차이는 없으나 감마분포에 의한 점근분포보다 2차 모형이 표준편차에서 상대적으로 큰 값을 가졌다. 특히 히스토그램에서 400mm 근처의 도수를 고려할 때 2차 연쇄-의존 모델의 적용이 적절하다 하겠다.

그림 4.4는 서산 지역의 봄의 46일간(1달 반)의 강수량에 대한 분포의 그림이다. 광주의 경우와 마찬가지로 감마분포에 의한 점근분포보다 2차 모형이 표준편차에서 상대적으로 큰 값을 가졌을 뿐만아니라 평균도 더 컸다. 여기서도 히스토그램에서 250mm 근처의 도수를 고려할 때 2차 연쇄-의존 모델의 적용이 적절하다고 보인다.

한편 기상학적인 측면에서 보면, 3개월간의 강수량의 분포는 기후변화에 따른 강수량

의 일정한 변동추세가 없다면 대체로 정규분포에 가까우나 (Wilks, 1995) 일부 강수량이 매우 많은 경우가 있다. 이는 특정 해, 즉 엘니뇨 또는 이상 몬순활동과 연관하여 강수량이 급격히 증가하는 해가 (간헐적으로) 존재함을 의미한다.

5. 요약 및 연구과제

강수 형태와 강수량을 동시에 모델링 하기 위해 사용되는 1차의 마코프 연쇄-의존 모델을 고차의 모델로 확장하였다. 또한 남한 53개 지역의 강수량 자료를 가지고 계절별 강수형태에 대한 마코프 연쇄의 차수를 선택하고 선택된 차수에 따른 추이확률을 계산하였다. 또한 차수 분포에 대한 기상학적인 의미를 기술하였다. 고차의 마코브 연쇄-의존 모델에 기초한 점근 정규분포를 이용하여 남한의 계절별 강수량의 분포 함수를 구할 수 있게 하였다.

지금까지 마코프 연쇄를 이용하여 강수 형태 및 강수량의 확률분포를 파악하였으나, 미래의 강수 가능성 및 강수량의 예측을 위해서는 강수에 영향을 주는 다른 많은 요인들(온도, 습도, 운량, 풍속, 기압 등)을 독립변수로 잡고 강수 형태 및 강수량을 종속변수로 하는 일종의 회귀모형이 필요하다(Stern and Coe, 1984). 나아가 이러한 회귀모형에 추가하여, 강수 형태는 인근 지역간에 유사성을 가지고 있기 때문에, 지역적 상관성을 고려하여 여러 지역을 한 모델에서 동시에 적합시키는 공간적 시간적 모형(Spatio-temporal model)의 수립이 요구되고 있다.

참고문헌

[1] 추춘기, 민경덕 (1983). 극동 아시아 지역의 저기압 활동에 관하여, 〈한국기상학회지〉, 19권 2호, 58-71.

[2] Guttorp, P. (1995). *Stochastic Modeling of Scientific Data*, Chapman & Hall.

[3] Katz, R.W. (1977). An Application of Chain-dependent Processes to Meteorology, *Journal of Applied Probability*, **14**, 598-603.

[4] Katz, R.W. (1981). On Some Criteria for Estimating the Order of a Markov Chain, *Technometrics*, **23**, 243-249.

[5] Kemeny, J., and Snell, J. (1983). *Finite Markov Chains*, Springer-Verlag, New York.

[6] Moon, S.E., Ryoo, S.B., and Kwon, J.G. (1994). A Markov Chain Model for Daily Precipitation Occurrence in South Korea, *International Journal of Climatology*, **14**, 1009-1016.

[7] O'Brien, G.L. (1974). Limit Theorem for Sums of Chain-dependent Processes, *Journal of Applied Probability*, **11**, 582-587.

[8] Ross, S. (1985). *Introduction to Probability Models*, Third edition, Academic Press, New York.

[9] Stern, R.D., and Coe, R. (1984). A Model Fitting Analysis of Daily Rainfall Data, *Journal of Royal Statistical Society A*, **147**, Part 1, 1-34.

[10] Suzuki, E. (1980). A Summarized Review of Theoretical Distributions Fitted to Climatic Factors and Markov Chain Models of Weather Sequences with Some Examples, *Statistical Climatology, Developments in Atmospheric Science*, **13**, 1-20.

[11] Wilks, D.S. (1995). *Statistical Methods in the Atmospheric Science*, Academic Press, New York.

[1999년 2월 접수, 1999년 6월 최종수정]

부록

[1차 마코브 연쇄-의존 모델에서 μ와 σ^2의 유도]

이 부분은 Guttorp(1995) 의 책에 간단히 나와 있으나 매우 불분명하여 이해하기 어려우므로 여기에 상세히 기술한다. 먼저 μ 의 유도는, $X_t = 1$ 인 경우만 생각하면 되므로, 다음과 같다.

$$\begin{aligned}\mu = E(Z_t) &= E[E(Z_t|X_{t-1}, X_t)] \\ &= \sum_i Z_t \left[\int P(Z_t|X_{t-1},\ X_t)dZ_t\right] P(X_{t-1},\ X_t) \\ &= \sum_i \mu_{i1} P(X_{t-1},\ X_t) \qquad (5.1) \\ &= \sum \mu_{i1}\ \pi_i\ p_{i1}. \qquad (5.2)\end{aligned}$$

식 (5.1)에서 $P(X_{t-1},\ X_t) = P(X_t = 1|X_{t-1} = i)P(X_{t-1} = i) = \pi_i\ p_{i1}$ 이다.

그리고 O'Brein(1974) 에 의해서 σ^2 은

$$\sigma^2 = \ \rho_0 + 2\sum_{j=1}^{\infty} \rho_j, \quad (0 < \sigma^2 < \infty) \qquad (5.3)$$

인데, 여기서 $\rho_j = Cov(Z_t, Z_{t+j})$, $j = 0, 1, 2\cdots$ 이므로,

$$\begin{aligned}\rho_0 = Var(Z_t) &= E[\ Var(Z_t|X_{t-1}, X_t)] + Var[\ E(Z_t|X_{t-1}, X_t)] \\ &= E(\sigma^2_{X_{t-1}}) + Var(\mu_{X_{t-1}}) \\ &= \sum \pi_i\ \sigma^2_{i1}\ p_{i1} + \sum \pi_i\ \mu^2_{i1}\ p_{i1} - \mu^2 \\ &= \pi_0 p_{01}(\sigma^2_{01} + \mu^2_{01}) + \pi_1 p_{11}(\sigma^2_{11} + \mu^2_{11}) - \mu^2. \qquad (5.4)\end{aligned}$$

여기서 $\sigma^2_{X_{t-1}}$와 $\mu_{X_{t-1}}$ 는 식 (3.1) 에서와 같은 의미로 Z_t 의 조건부 분산 및 조건부 평균이다. 그리고

$$\rho_j = E(Z_t Z_{t+j}) - E(Z_t)E(Z_{t+j}) \tag{5.5}$$

이므로, 두번째 항은 식 (5.2)로부터 구해지고, 첫번째 항은

$$E(Z_t Z_{t+j}) = E[\, E(Z_t\, Z_{t+j}|\, X_{t-1}, X_t, X_{t+j-1}, X_{t+j})\,] \tag{5.6}$$

을 이용한다. 식 (5.6)를 구하기 위해 먼저 다음을 구한다.

$$\begin{aligned}
P(\underline{X}) &= P(X_{t-1}=i, X_t=1, X_{t+j-1}=k, X_{t+j}=1) \\
&= P(X_{t+j}=1|X_{t+j-1}=k, X_{t-1}=i, X_t=1) \\
&\quad \cdot P(X_{t+j-1}=k|X_t=1, X_{t-1}=i)P(X_t=1|X_{t-1}=i)P(X_{t-1}=i) \\
&= P(X_{t+j}=1|X_{t+j-1}=k)P(X_{t+j-1}=k|X_t=1) \\
&\quad \cdot P(X_t=1|X_{t-1}=i)P(X_{t-1}=i) \\
&= p_{k1}\, p_{1k}^{j-1}\, p_{i1}\, \pi_i.
\end{aligned}$$

이를 이용하여 식 (5.6)은 다음과 같이 된다.

$$\begin{aligned}
\text{식 (5.6)} &= \sum_{i\,k}\left[\int\int Z_t Z_{t+j} P(Z_t Z_{t+j}|\underline{X}) dZ_t dZ_{t+j}\right] P(\underline{X}) \\
&= \sum_{i\,k}\left[\int Z_t P(Z_t|\underline{X})dZ_t\right]\left[\int Z_{t+j} P(Z_{t+j}|\underline{X})dZ_{t+j}\right] P(\underline{X}) \\
&= \sum_{i\,k} \mu_{i1}\, \mu_{k1}\, p_{k1}\, p_{1k}^{j-1}\, p_{i1}\, \pi_i.
\end{aligned} \tag{5.7}$$

또 식 (5.3)에서의 $\sum_{j=1}^{\infty}\rho_j$ 를 구하기 위해 ρ_j 를 식 (5.5)로부터 식 (5.2)와 식 (5.7)을 이용하여 행렬로 표현하면 다음과 같다.

$$\begin{aligned}
\rho_j &= \pi Q P^{j-1} Qe - (\pi Qe)(\pi Qe) \\
&= \pi Q(P^{j-1} - \Pi)Qe,\ j = 1, 2, \cdots .
\end{aligned} \tag{5.8}$$

여기서 $\pi = (\pi_0\ \pi_1)$, $\Pi = e\pi$, $e^t = (1\ 1)$, $P^0 = I$, $Q = \{p_{ij}\ \mu_{ij}\}$ 이다.

이제 식 (5.8) 으로부터 $\sum_{j=1}^{\infty}(P^{j-1} - \Pi)$ 을 구해보자. 그런데 Kemeny and Snell(1983)의 Theorem 4.3.1에 의해서 다음을 얻는다.

$$\sum_{j=1}^{\infty}(P^{j-1} - \Pi) = \ [I - (P - \Pi)]^{-1} - \Pi.$$

여기서 $[I - (P - \Pi)]^{-1}$은 마코프 연쇄의 Fundamental 행렬이라고 불리며 이 행렬은 항상 존재한다(Kemeny and Snell, 1983). 따라서 결국 σ^2은 식 (5.3)으로부터 (ρ_0 는 식 (5.4)로 표현되고) 다음과 같이 쓰여진다.

$$\sigma^2 = \rho_0 + 2\pi Q\ ([I - (P - \Pi)]^{-1} - \Pi)\ Qe.$$

Analysis of Daily Precipitation in South Korea Using a Higher Order Markov Chain-dependent Model*

Jeong-Soo Park[1] Young-Geun Chung[2] Rae-Seon Kim[3]

ABSTRACT

A Markov chain-dependent model is commonly used to fit daily precipitation occurrence and its amounts simultaneously. Its extension to a higher order Markov chain-dependent model is presented. The model is applied to daily precipitation data of 53 stations for 15 years in South Korea. The order of Markov Chain(determined by BIC) and asymptotic distribution of seasonal precipitation are obtained. Spatial and seasonal patterns of precipitation are described.

* This study was financially supported by Chonnam National University in the program, 1998.

1) Associate Professor, Department of Statistics, and Institute of Basic Sciences, Chonnam National University
2) Professor, Department of Earth Science Education, Chonnam National University
3) Graduate, Department of Statistics, Chonnam National University