응용통계연구 제 19권 3호, 2006년, pp. 431–445

기온 강수량 자료의 함수적 데이터 분석*

강기훈[1] 안홍세[2]

요 약

본 연구는 함수적 데이터 분석의 몇 가지 이론에 대해 소개하고 분석 기법을 실제 자료에 적용하는 내용을 다루었다. 함수적 데이터 분석의 이론적 내용으로 기저를 이용해 자료를 함수적 데이터로 표현하는 방법, 그리고 함수적 데이터의 변동성을 조사하는 주성분분석, 선형모형 등에 대해 살펴보았다. 그리고 우리나라 기온 데이터와 강수량 데이터를 대상으로 각각 함수적 데이터 분석 기법을 적용해 보았다. 또한, 기온과 강수량 데이터에 대해 함수적 회귀모형을 적합시켜 두 변수간의 함수관계를 살펴보았다.

주요용어: 함수적 주성분분석, 함수적 분산분석, 함수적 선형모형

1. 서론

현실 세계에 존재하는 다양한 분야의 데이터는 양이나 질의 점진적인 변화를 나타내는 임의의 곡선으로 묘사되어 표현되어질 수 있고, 통계학의 적용분야가 다양해짐에 따라 이렇게 관측치가 곡선(curves) 또는 영상(images)의 형태가 되는 경우가 흔해졌다. 이런 곡선이나 영상을 함수(function)라고 할 수 있으며, 관측의 정도는 각 점들에서 선의 조각 또는 면이나 부피의 비율로써 나타낼 수 있다. 시간에 따른 기온의 변화 데이터가 그 대표적인 예라고 할 수 있는데 이런 함수는 연속적인 성질에도 불구하고 관측의 속성상 이산형으로 나타난다. 이런 이유로 관측된 곡선들과 영상들을 함수적 데이터(functional data)라고 하며 이 데이터를 분석하는 통계적 분야를 함수적 데이터 분석(functional data analysis : FDA)이라고 한다. 이런 이산형 관측치를 가지고 모함수 형태를 근사시키기 위한 방법들에 대해서는 이미 많은 연구가 이루어져 있다. 예를 들면, 유한한 공간에서 보간법(interpolation), 평활법(smoothing)을 이용하거나 기저의 선형결합으로 함수를 표현하는 방법 등이 있다.

함수적 데이터 분석 기법은 Ramsay와 Dalzell (1991)에 의해 일반화되었고, 세부적인 내용에 대해서는 Locantore 등 (1999), Ramsay와 Silverman (2002, 2005) 등을 참고하기 바란다. 함수적 데이터 분석의 목적은 여타 다른 분야의 통계적 분석의 목적과 같이 표본이 뽑힌 모집단에 대한 추론을 하는 것이다. 이를 위해 데이터를 변환 또는 재표현하여 추후 분석을 용이하게 하거나 또는 자료의 특성을 잘 묘사할 수 있는 그림을 제공한다든가, 자료

* 본 연구는 과학기술부/한국과학재단 우수연구센터육성사업의 지원으로 수행되었음 (R11-2000-073-00000)
1) (449-791) 경기도 용인시 처인구 모현면 한국외국어대학교 정보통계학과, 부교수
E-mail: khkang@hufs.ac.kr
2) (130-791) 서울시 동대문구 이문동 한국외국어대학교 대학원 통계학과, 석사
E-mail: babaman@empal.com

의 패턴 또는 변동의 중요한 요인을 파악한다. 또한 함수적 데이터로 이루어진 입력변수들의 정보를 이용하여 반응변수의 변동성을 설명하기도 한다.

함수적 데이터 분석에서 자료의 변동성을 알아보기 위해서는 정교한 방법들이 필요하며 이런 방법들에는 함수적 주성분분석(functional principal components analysis), 함수적 분산분석(functional analysis of variance), 함수적 선형모형(functional linear model) 그리고 함수적 정준상관분석(functional canonical correlation) 등이 있다. 또한, 함수적 데이터 분석에서는 변동성의 설명 등을 위해 함수의 미분계수와 선형 미분연산자(linear differential operators)들이 유용하게 사용된다.

본 연구에서는 함수적 데이터 분석에서 이러한 방법론들에 대해 간단히 소개하고, 실제 자료에 적용시킴으로써 함수적 데이터 분석 기법의 유용성에 대해 살펴볼 것이다. 이를 위해 전형적인 함수적 데이터라 할 수 있는 기온 및 강수량 데이터를 이용할 것이다. 지금까지 함수적 데이터 분석 방법을 이용해 우리나라 기온 및 강수량에 대해서 분석한 것은 없었으며 생활과 밀접한 이들 데이터에 대해서 분석 기법을 적용해 보는 것은 의미있는 일이라고 할 수 있다. 본 논문에서는 1970년 1월부터 2004년 5월까지 우리나라에서 조사된 기온 데이터와 같은 기간에 조사된 강수량 데이터를 이용하였다.

본 논문의 구성은 다음과 같다. 2장에서는 함수적 데이터를 분석하는 방법 중에 몇 가지를 간단히 소개한다. 3장에서는 실제 우리나라 기온과 강수량 데이터를 가지고 함수적 데이터 분석 기법을 적용한 결과에 대해 살펴본다. 4장에서는 실제 함수적 데이터 분석의 결과를 통해 확인된 내용과 함수적 데이터 분석의 추후 과제에 대해서 언급할 것이다.

2. 함수적 데이터 분석 기법

함수적 데이터 분석의 가장 기본적인 원리는 데이터가 개별적인 관측치의 연속체가 아니라 하나의 형태를 가지고 관측된 함수라는 것이다. 그러나 실제로 함수적 데이터는 이산적인 형태로 관측되거나 측정된다. 함수적 관측치는 n개의 쌍 (t_j, y_j)로 구성되며, y_j는 $t = t_j$에서 함수값 $x(t)$의 관측이나 기록이다. 이렇게 이산형 형태로 관측된 데이터에는 대부분 오차가 존재하며, 가공되지 않은 데이터를 함수적 데이터로 표현하기 위해서는 다양한 평활화(smoothing) 기법들을 필요로 한다. 이 절에서는 함수적 데이터 분석에 관련된 몇 가지 기법들에 대해 살펴보고자 한다.

2.1. 기저 함수 방법(basis function methods)

데이터를 평활화(smoothing)하는 것과 유사한 함수적 데이터 분석은 이산적인 관측치 $\{(t_j, y_j) | j = 1, \cdots, n\}$가 주어졌을 때, 모형

$$y_j = x(t_j) + \varepsilon_j \tag{2.1}$$

를 사용하여 모함수 $x(t)$를 추정하고자 하는 것이다. 이 때, 평활법과 가장 유사한 방법 중에 하나는 알려진 K개의 기저 함수 ϕ_k의 선형 결합으로 함수를 표현하고자 하는 방법이다.

$$x(t) = \sum_{k=1}^{K} c_k \phi_k(t) \tag{2.2}$$

이 경우 $x(t)$의 평활 정도는 기저 함수의 수 K에 의해서 결정되는데, 큰 값의 K는 평활정도가 작아 데이터를 거의 따라 움직이게 하므로 편의(bias)를 줄이고 작은 값의 K는 평활정도가 커서 분산을 작게하는 경향이 있어 커널형 함수추정에서 평활량(bandwidth)의 선택과 유사하면서 그 크기의 영향은 상반된다. K의 선택과 관련된 자세한 사항은 Ramsay와 Silverman (2005)의 4.5절을 참고하기 바란다.

식 (2.2)에서 상수c_k는 오차제곱합인 식 (2.3)을 최소로 하는 최소제곱법을 적용시켜 얻을 수 있다.

$$SMSSE = \sum_{j=1}^{n}[y_j - \sum_{k=1}^{K} c_k\phi_k(t_j)]^2. \tag{2.3}$$

여기서, $\mathbf{y} = (y_1, \ldots, y_n)'$, $\mathbf{c} = (c_1, \ldots, c_K)'$ 그리고 $j = 1, \ldots, n, k = 1, \ldots, K$에 대해 (j,k) 원소를 $\phi_k(t_j)$로 갖는 행렬을 $\mathbf{\Phi}$라 하면 식(2.3)은 다음과 같이 행렬로 표현된다.

$$SMSSE(\mathbf{y}|\mathbf{c}) = (\mathbf{y} - \mathbf{\Phi c})'(\mathbf{y} - \mathbf{\Phi c}) = ||\mathbf{y} - \mathbf{\Phi c}||^2 \tag{2.4}$$

식 (2.4)를 최소로 하는 $\hat{\mathbf{c}}$는 $\hat{\mathbf{c}} = (\mathbf{\Phi}'\mathbf{\Phi})^{-1}\mathbf{\Phi}'\mathbf{y}$로 구해짐은 쉽게 확인할 수 있다.

식 (2.2)의 표현을 위해 기저 함수를 선택할 때 바람직한 기준은 모함수와 추정된 함수가 비슷한 형태를 가지도록 하는 것이다. 물론, 비교적 적은 수의 기저를 사용하면서 잘 근사하도록 선택하는 것이 계산을 용이하게 하며 좋은 추정 결과를 낳을 것이다. 일반적으로 잘 알려진 퓨리에 기저를 이용한 표현은 식 (2.2)의 표현에서 기저를 $\phi_0(t) = 1$, $\phi_{2k-1}(t) = \sin kwt$ 그리고 $\phi_{2k}(t) = \cos kwt$로 정의하면 다음과 같이 나타낼 수 있다.

$$x(t) = c_0 + c_1 \sin wt + c_2 \cos wt + c_3 \sin 2wt + c_4 \cos 2wt + \cdots$$

퓨리에 기저는 주기적이고 주기는 $2\pi/w$에 의해 결정되며 지역적 성질이 적고 곡률의 차수가 동일한, 즉, 급격한 변화가 없는 함수에 대해 특히 유용하게 적용되며 본 논문에서도 사용될 것이다. 이 밖에 스플라인(spline)이나 웨이브릿(wavelets) 등 다른 기저함수의 사용과 이와 관련된 노트(knot)나 차수(order)의 선택에 대해서는 Ramsay와 Silverman (2005)의 3.5, 3.6절을 참고하면 된다.

2.2. 함수적 주성분분석 (functional principal component analysis)

복잡한 함수적 데이터로부터 핵심적이고 전형적인 형태를 얻는 방법인 함수적 주성분분석에 대해 개략적으로 설명하고자 한다. 함수적 데이터 분석에서 주성분 요소들은 물론 함수이며, 그 형태나 구조는 함수의 증감부분, 최대최소 또는 국소적 성질 등에 의해 결정된다. 그러므로, 자료로부터 구해지는 주성분 요소들에 대해서는 보다 면밀하게 살펴볼 필요가 있다. 첫 번째 주성분은 자료 변동의 가장 많은 부분을 설명하며, 두 번째 주성분은 첫 번째 주성분과 직교하는 변동을 설명하는 식으로 몇 개의 주성분으로 데이터 변동의 거의 전부를 설명할 수 있다.

간편한 표현을 위해 관측된 데이터의 함수적 표현인 $x_i(t), i = 1, \cdots, n$이 $t \in T$인 모든 t에 대해 $\sum x_i(t) = 0$을 만족한다고 하면(표준화로 가능) 함수 공간에서 표본에 기초한 공

분산 함수는 다음과 같이 정의된다.

$$cov_X(s,t) = v(s,t) = n^{-1}\sum_{i=1}^{n} x_i(s)x_i(t),\ s,t \in T.$$

함수적 데이터 $x(s)$의 공분산 함수는 적분변환을 사용하여 고유값 θ_j와 각각의 고유값에 대응되는 고유함수(eigenfunction) ψ_j로 나타낼 수 있다. 주성분 함수를 찾기 위해서는 고유치 θ에 대해서 다음의 고유방정식(eigenequation)을 만족하는 해를 찾으면 된다. 이 경우에 고유치의 크기에 따라 주성분 함수의 순서가 정해지는 것이다.

$$\int v(s,t)\psi(t)dt = \theta\psi(s). \tag{2.5}$$

식 (2.5)는 $V\psi = \int v(\cdot,t)\psi(t)dt$라 하면 $V\psi = \theta\psi$로 나타내어질 수 있고 이는 적절한 표기법을 사용하면 다변량 분석에서 주성분분석을 위한 표기법과 일치함을 알 수 있다. 또한, 다변량 분석에서 주성분 벡터와 마찬가지로 주성분 함수들에 대해 $||\psi||^2 = 1, <\psi_k, \psi_m> = 0 (k < m)$이 요구된다. 참고로, 다변량 자료에서 주성분분석과의 비교는 Ramsay와 Silverman (2005)의 8.2절에 제시되어 있다.

2.3. 함수적 분산분석(functional analysis of variance)

주어진 함수적 데이터를 몇 개의 그룹에서 관측한 것으로 나누고 그룹별로 유의한 차이가 있는지를 확인하는 것은 분산분석의 문제에 해당된다. 종속변수는 함수적 데이터의 관측치인 $y(t)$이고 그룹별 비교는 분산분석의 형태이기 때문에 방법론적으로 함수적 분산분석(FANOVA)이라고 할 수 있다. 고려하는 그룹의 수가 g개일 때, j 그룹에서 i번째 관측치를 y_{ij}라 하면 모형을 다음과 같이 나타낼 수 있다.

$$y_{ij}(t) = \mu(t) + \alpha_j(t) + \varepsilon_{ij}(t),\ i = 1,\cdots,k,\ j = 1,\cdots,g. \tag{2.6}$$

여기서 함수 $\mu(t)$는 전체 평균 함수를 나타내며 α_j항은 j번째 그룹의 효과를 나타내는 함수에 해당된다. 또한, 오차 함수 $\varepsilon_{ij}(t)$는 설명되지 않는 변동성을 나타낸다. 이런 분산분석을 위해서는 모든 t에 대해 $\sum_j \alpha_j(t) = 0$ 이라는 제약 조건이 필요하다. 이 때, 각 행의 첫 번째 열과 $(j+1)$번째 열이 1이고 나머지는 0인 $kg \times (g+1)$인 계획행렬(design matrix) $\mathbf{Z}$를 정의하고, $\beta_1 = \mu, \beta_2 = \alpha_1, \cdots, \beta_{g+1} = \alpha_g$를 원소로 갖는 함수적 벡터를 β라 하면 식 (2.6)을 다음과 같이 행렬식으로 표현할 수 있다.

$$y = \mathbf{Z}\beta + \varepsilon. \tag{2.7}$$

여기서 β는 오차 제곱합 $||y - \mathbf{Z}\beta||^2$을 최소로 만드는 최소제곱 추정치 $\hat{\beta}$로 추정할 수 있다.

추정된 함수적 분산분석 모형의 평가를 위해서는 일반적인 분산분석에서 사용되는 오차제곱합인 SSE, 총 편차제곱합인 SSY, 다중상관계수의 제곱 R-square(RSQ)와 평균제곱의 비율인 F-ratio 함수를 사용할 수 있는데 다만 이러한 항들이 함수라는 것만 차이가 있

으며 다음과 같이 나타낼 수 있다.

$$\begin{aligned} SSE(t) &= \sum_{i,j}[y_{ij}(t) - (\mathbf{Z}\hat{\beta})_{ij}(t)]^2 \\ SSY(t) &= \sum_{i,j}[y_{ij}(t) - \hat{\mu}(t)]^2 \\ RSQ(t) &= [SSY(t) - SSE(t)]/SSY(t) \\ MSE(t) &= SSE(t)/df(error) \end{aligned}$$

여기서 $df(error)$는 잔차의 자유도를 나타내는 것으로 $(g-1)\times k$ 임을 알 수 있으며, 모형의 자유도는 $df(model) = (g-1)$이므로

$$MSR(t) = \frac{SSY(t) - SSE(t)}{df(model)}$$

이고, 모형의 유의성 검정을 위한 F-ratio함수는 아래와 같이 계산할 수 있다.

$$FRATIO(t) = \frac{MSR(t)}{MSE(t)} \tag{2.8}$$

식 (2.8)의 통계량은 자유도가 $(g-1, kg-g)$인 F분포를 따르므로 이 값이 유의수준 α의 기각치보다 크면 모형이 유의하다고 판단한다.

3. 함수적 데이터 분석 기법의 적용

3.1. 기온 데이터(temperature data)

그림 3.1은 우리나라 68개 기상 관측소에서 1970년 1월부터 2004년 5월까지 측정된 기온을 나타내는데 작은 원으로 표시된 각각의 점들은 매월 각 기상 관측소에서 약 30년 동안 기록된 평균기온을 나타낸다. 서로 다른 각 선들은 기상관측소들을 구분하여 나타낸다. (관측 지점: 서울, 인천, 수원, 강화, 양평, 이천, 철원, 춘천, 원주, 인제, 홍천, 속초, 대관령, 강릉, 울릉도, 태백, 충주, 청주, 추풍령, 제천, 보은, 서산, 대전, 천안, 보령, 부여, 금산, 군산, 전주, 부안, 임실, 정읍, 남원, 장수, 광주, 목포, 여수, 완도, 순천, 장흥, 해남, 고흥, 울진, 안동, 포항, 대구, 봉화 ,영주, 문경, 영덕, 의성, 구미, 영천, 울산, 마산, 부산, 통영, 진주, 거창, 합천, 밀양, 산청, 거제, 남해, 제주, 고산, 서귀포, 성산포)

이 데이터는 주기가 12라고 할 수 있고 따라서 12개의 퓨리에 기저와 상수항을 이용하여 주어진 이산형 관측치들을 함수적 데이터로 변환하여 부드러운 68개의 곡선으로 나타내면 그림 3.2와 같다. 대체적으로 월별 평균 기온의 변화는 사인 함수와 비슷한 형태로 보이고 있다. 본 논문에서 사용된 프로그램은 함수적 데이터 분석과 관련된 홈페이지 (http://www.functionaldata.org)에서 구할 수 있고, S-plus 버전을 우리의 목적에 맞게 변경하여 사용하였다.

그림 3.3은 우리나라 기온의 평균과 분산을 나타낸 것이다. 이미 예상하듯이 평균 기온은 여름철에 높고 겨울철에 낮으며 봄·가을에는 서늘한 기온 형태로 사계절이 뚜렷하게 나

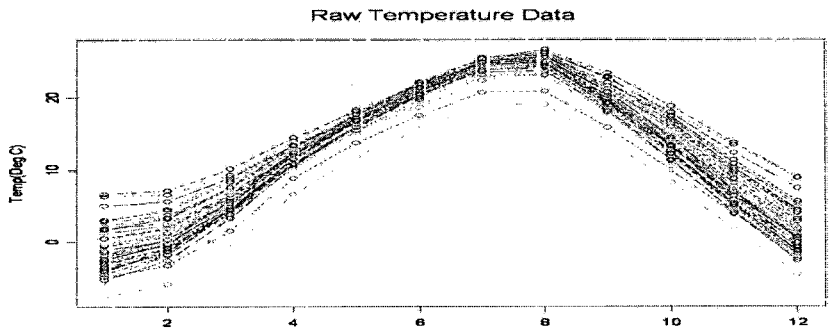

그림 3.1: 우리나라 68개 기온관측소에서 관측한 월별 평균 기온

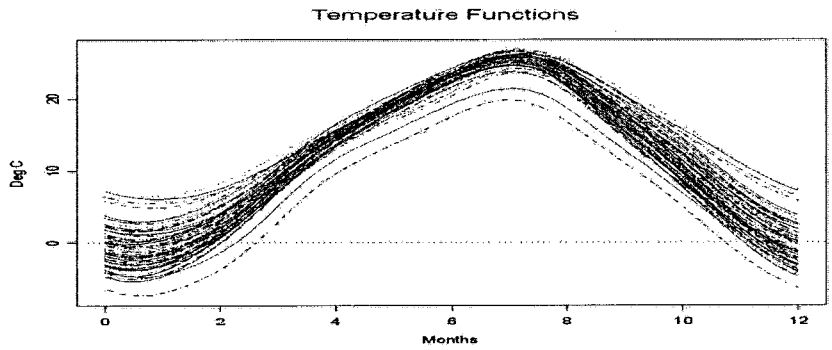

그림 3.2: 그림 3.1을 함수적 데이터로 표현한 곡선

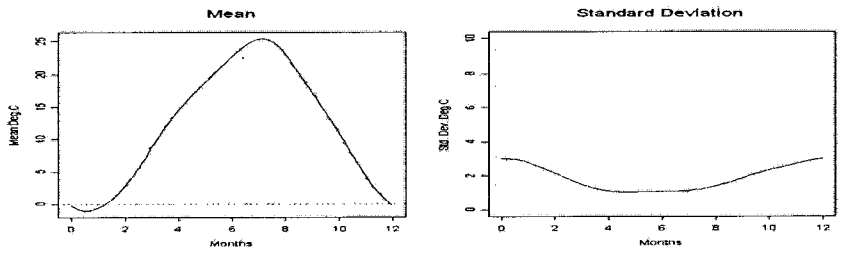

그림 3.3: 함수적 데이터로 표현한 기온의 평균과 분산함수

타내고 있다. 그리고 분산의 경우는 여름철이 겨울철보다 더 작다는 것, 즉 겨울철에 지역별로 기온 차이가 더 크게 나타남을 알 수 있다.

전통적인 주성분분석은 다차원적인 변수들을 축소, 요약하는 차원의 단순화와 더불어 일반적으로 서로 상관되어 있는 반응변수들 간의 복잡한 구조를 분석하는데 그 목적을 두고 있다. 이를 위하여 주성분분석은 반응변수들을 선형변환시켜, 주성분이라고 부르는 서로 상관되어 있지 않은, 혹은 독립적인 새로운 인공 변수들을 유도한다. 함수적 주성분분석 또한 데이터 변동성의 형태를 찾기 위해 계산이 가능한 유한한 주성분 요소를 찾음으로써 함수적 데이터의 차원을 축소한다. 함수적 주성분 요소의 적재는 역시 함수이고, 때때로 주성분 요소의 평활화(smoothing)가 필요하다. 2.2절에서 언급한대로 함수적 주성분분석에 대한 개념을 다변량 자료에서 주성분분석과 연관시켜 설명하는 내용은 Ramsay와 Silverman (2005)의 8.2절을 참고하고, 구체적인 계산 방법은 8.4절에 자세하게 언급되어 있으므로 여기서는 생략하기로 한다.

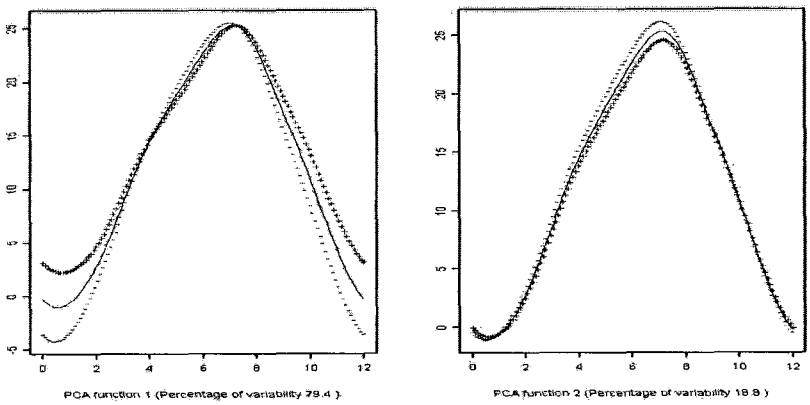

그림 3.4: 기온 평균 곡선에서 각 주성분 곡선의 배수를 더한(+)효과와 뺀(-)효과

함수적 주성분 분석의 결과를 확인하기 가장 좋은 방법이 그림으로 표현하는 것이다. 그림 3.4는 기온에 대한 주성분 분석의 결과를 나타내고 있다. 실선은 전체 평균 함수를 나타내며 각 주성분 곡선을 더한 효과와 뺀 효과를 나타낸 그림이다. 첫 번째 주성분은 전체 변동성의 79.4%를 설명해 주고 있으며 우리나라 기온의 11월 말 부터 다음 해 1월 말까지의 겨울철 변동성을 설명해 주고 있다. 두 번째 주성분은 전체 변동성의 18.8%를 설명해 주고 있으며 여름철의 변동성을 설명해 주고 있다.

3.2. 강수량 데이터(precipitation data)

한 해 동안 우리나라의 강수량은 곳에 따라 많은 차이가 난다. 강수량이란 비 뿐만 아니라 눈, 우박, 서리, 이슬 등 땅 위에 내린 물의 양을 모두 더한 것을 말한다. 물론 우리나라 강수량의 대부분은 비가 차지한다. 우리나라에서 강수량은 남쪽에서 북쪽으로 올라갈수록

줄어들고, 바람의 방향, 해안과의 거리, 지형의 영향을 크게 받는다(정창희, 1976).

우리나라 68개 기상 관측소로부터 측정된 강수량 데이터를 살펴보면 연간 강수량의 변화에 대해서 알 수 있다. 그림 3.5는 강수량의 이산형 관측치들을 함수적 데이터로 변환하여 부드러운 곡선으로 나타낸 것이다. 여기서 각 곡선의 모양은 68개 지점(기온 데이터와 동일)의 기상 관측소들을 구분하여 나타낸다. 그림 3.5를 보면 지역별로 강수량의 차이가 크다는 것과 여름철에 강수량이 집중됨을 알 수 있다.

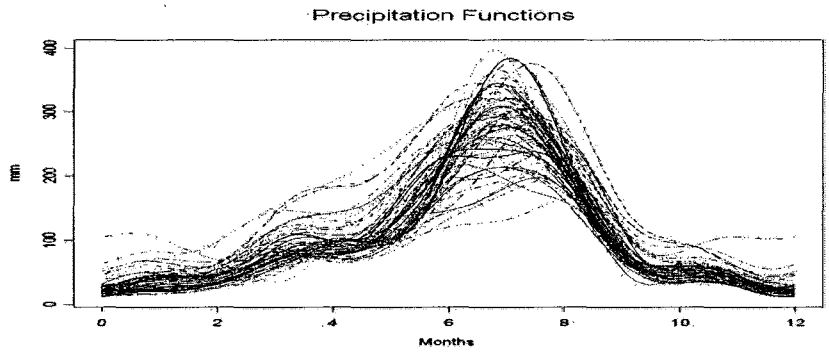

그림 3.5: 월별 평균 강수량의 함수적 데이터 표현

그림 3.6은 6개의 특징 있는 기상 관측소에서 평균 강수량을 나타낸 그림이다. 여름철에 가장 강수량이 많은 지역은 철원 지역이고, 두 번째로 강수량이 많은 지역은 서울 지역인데 두 지역 모두 여름철 이외의 지방에서는 평균 강수량보다 적게 나타나고 있다. 울릉도 지역은 다른 지역과 달리 여름철에 강수량이 집중된 것이 아니라 연중 고르게 분포되게 나타나고 있다. 이렇게 강수량은 지역별로 다른 형태로 나타나고 있다.

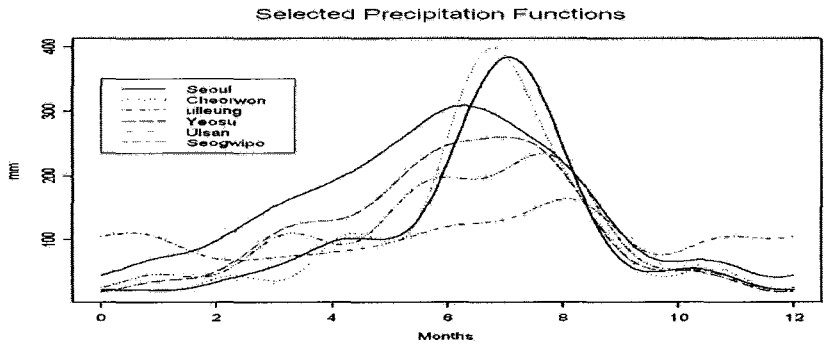

그림 3.6: 선택된 6개 지역의 강수량: 7월에 높은 순으로 철원, 서울, 서귀포, 여수, 울산, 울릉도

그림 3.7은 우리나라 강수량의 평균과 분산을 나타낸 것인데 여름철이 강수량의 대부분

을 차지하며 겨울철에는 강수량이 적음을 알 수 있다. 강수량의 분산을 보면 여름철에 크고 겨울철에 작은 형태를 보이고 있다. 이것은 여름철에는 어느 지역이나 강수량이 많으나 지역별로 큰 차이가 있기 때문이라고 할 수 있으며 겨울철에는 어느 지역이나 비가 적게 내리기 때문이다.

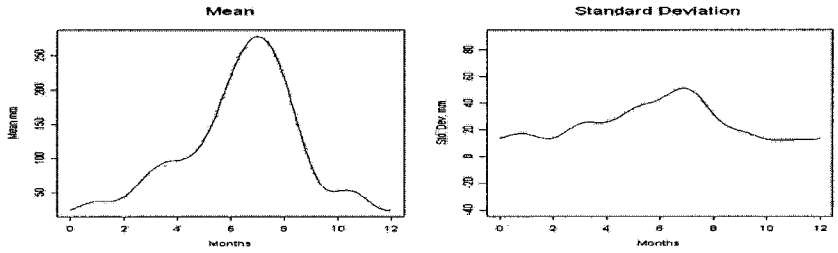

그림 3.7: 강수량 데이터의 평균 함수와 분산 함수

강수량 데이터도 기온의 경우와 유사하게 주성분분석을 시행하였다. 그림 3.8은 강수량에 대한 주성분분석의 결과이다. 실선으로 전체 평균 강수량이 각각의 그림에 그려져 있으며 또한 각각의 평균 곡선에 각 주성분 곡선의 배수를 더하거나 뺀 것을 함께 나타내었다. 첫 번째 주성분은 전체 변동성의 43.6%를 설명해 주고 있으며, 여름철 강수량의 변동성을 설명해 주고 있다. 두 번째 회전된 주성분은 전체 변동성의 39.5%를 설명해 주고 있으며, 봄과 초여름의 변동성을 설명해 주고 있다. 그리고 세 번째 주성분은 전체 변동성의 16.6%의 변동성을 설명해 주고 있으며 겨울철의 변동성을 설명해 주고 있다.

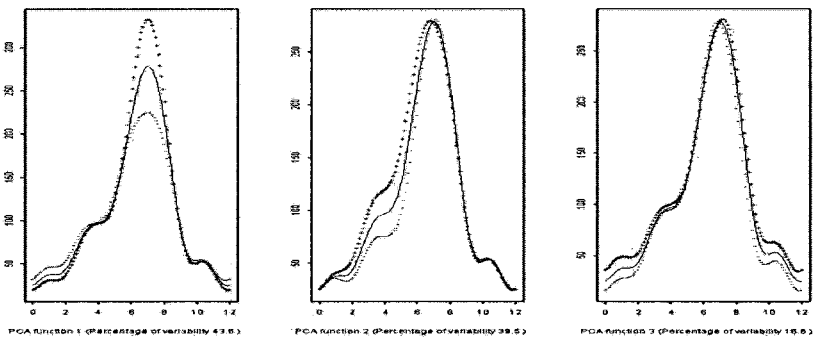

그림 3.8: 평균 강수량과 주성분 효과를 더하거나(+) 뺀(-) 효과

3.3. 기온 강수량 자료의 함수적 분산분석

문영수(1996)는 우리나라 전체를 일조율에 의해 태백산맥 이서의 중서부지역(Part 1), 지리산계의 서남지역(Part 2), 영동지역과 영남북부지역(Part 3), 영남지역(Part 4), 울릉도 지역(Part 5), 제주도지역(Part 6)의 6개 권역으로 나누었다. 기온 데이터의 경우에도 이 구분이 의미가 있는 것인지 확인하기 위하여 함수적 분산분석 기법을 적용하여 보았다. 그림 3.9는 모형 (2.6)을 적용하여 6개 지역에서 측정된 기온의 지역별 영향인 $\alpha_j, j = 1, \cdots, 6$의 추정값을 나타낸 것이며 그림 3.10은 $\mu + \alpha_j$의 추정치를 우리나라 전체 기온의 평균 함수 μ와 함께 나타낸 그림이다.

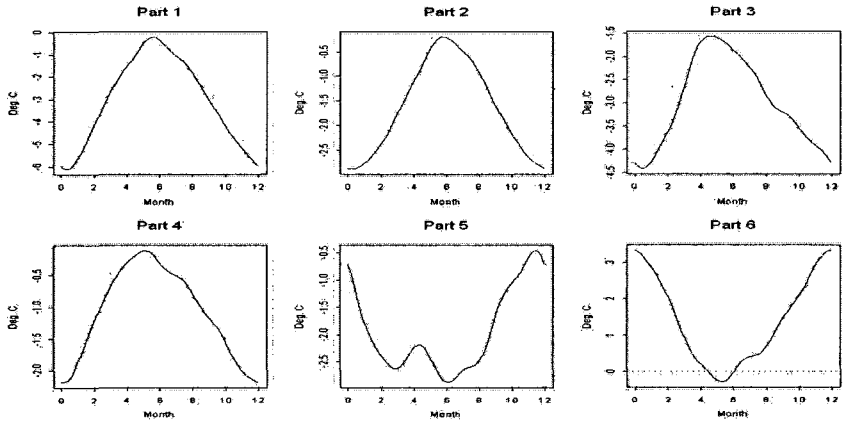

그림 3.9: 모형 $Temp_{ij}(t) = \mu(t) + \alpha_j(t) + \varepsilon_{ij}(t)$ 에서 기온 함수의 지역적 영향 α_j의 추정치

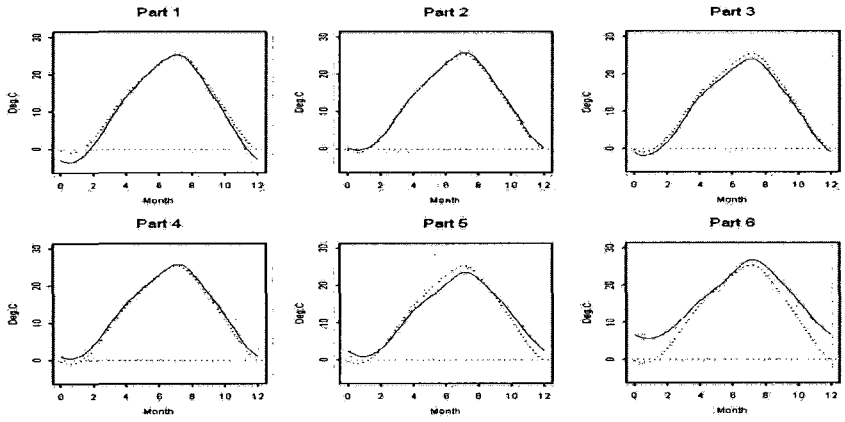

그림 3.10: 평균 기온 함수 μ(점선)와 각 지역에서 추정된 기온 함수 $\mu + \alpha_j$(실선)

Part 1 지역에서 봄 기온은 우리나라의 평균 기온과 유사하지만 다른 계절에는 낮게 나타나고 있으며 특히 겨울에 큰 차이를 보이고 있다. Part 2 지역에서는 봄철에 우리나라의 평균 기온과 유사하게 나타나고 있지만 다른 계절에서는 약간 높은 온도를 보이고 있다. Part 3 지역은 전 계절에서 우리나라의 평균기온보다 낮게 나타나고 있다. Part 4 지역은 영남 내륙 분지와 남동해안지역을 포함하는 지역으로서, 우리나라의 평균기온보다 전 계절에서 높게 나타나고 있으며 특히 겨울에 많은 차이를 보이고 있다. Part 5 지역은 동해안에서 머리 떨어져 해양성 기후를 띠는 지역으로, 우리나라의 평균 기온보다 여름에는 낮게 그리고 겨울에는 높게 나타나고 있다. Part 6 지역에서는 봄, 초여름에서는 우리나라의 평균 기온보다 낮지만 늦가을과 겨울에는 상당히 높은 형태를 보이고 있다.

그림 3.11는 다중상관계수의 제곱인 RSQ와 식 (2.8)에서 정의된 F-ratio함수를 나타내고 있다. 그림 3.11에서 RSQ는 상대적으로 낮게 나타나는 시점도 있지만 전체적으로 그리 낮은 것은 아니며 F-ratio는 5% 유의수준에서 기준이 되는 기각치 2.36보다 높게 나타났다. 그러므로 6개 지역(Part 1 - Part 6)에서 지역별로 기온의 차이가 유의하다고 할 수 있다.

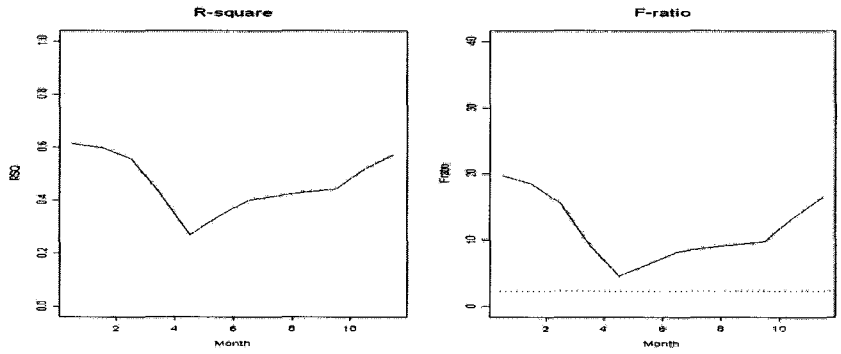

그림 3.11: RSQ와 F-ratio 함수. 오른쪽에서 수평 점선은 자유도가 (5, 62)인 F-분포의 유의수준 5% 기각치

기온 데이터와 동일하게 적용된 지역구분을 바탕으로 강수량 데이터에 적용하여 우리나라 평균 기온함수 $\mu(t)$와 6개 지역의 영향이 추가된 $\mu(t)+\alpha_j(t)$의 추정 결과를 그림 3.12에 나타내었다. 이에 대한 해석은 기온의 경우와 비슷하게 할 수 있으며, RSQ와 F-ratio도 모두 유의하게 나타나므로 결과는 생략하였다.

3.4. 함수적 반응변수와 함수적 공변량의 선형모형

반응변수와 공변량이 모두 함수인 경우에 함수적 선형모형의 적합에 대해 강수량과 기온 데이터로 살펴보자. 예를 들어, 우리나라에서 기온자료 Temp의 정보로부터 강수량 Prec의 추정에 대해 관심이 있다고 하자. 기온과 강수량은 모두 로그 변환한 것을 사용하기로 한다. 이 경우에 관심이 되는 모형은 다음과 같이 생각해 볼 수 있다.

$$\mathrm{Prec}_i(t) = \alpha(t) + \mathrm{Temp}_i(t)\beta(t) + \varepsilon_i(t), \;\; i = 1, \cdots, N. \tag{3.1}$$

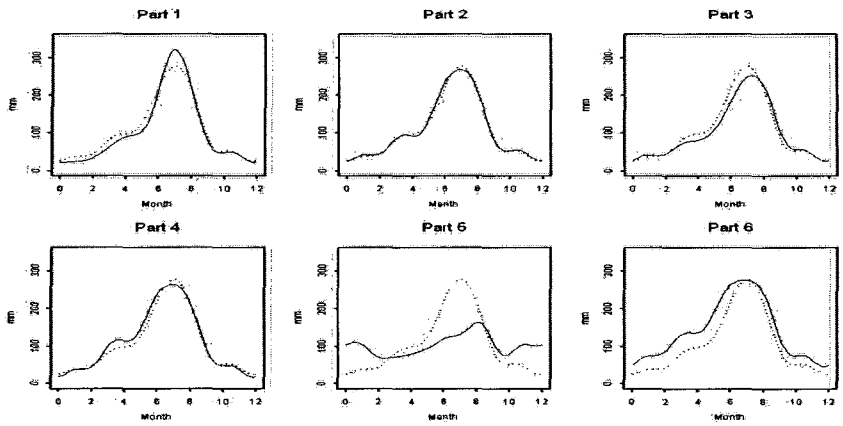

그림 3.12: 평균 강수량 함수 μ(점선)와 각 지역에서 추정된 강수량 함수 $\mu+\alpha_j$(실선)

모형 (3.1)은 t 시점에서 강수량에 t 시점에서만의 기온의 영향을 고려한 것이다. 또한, 고려할 수 있는 다른 모형으로는

$$\mathrm{Prec}_i(t) = \alpha(t) + \int \mathrm{Temp}_i(s)\beta(s,t)ds + \varepsilon_i(t), \quad i = 1, \cdots, N. \tag{3.2}$$

가 있는데, 이는 일년 또는 일정 기간의 기온이 강수량에의 영향을 고려하는 것에 해당된다. 모형 적합의 기준으로는 함수적 버전의 오차제곱합에 해당되는 다음 적분된 제곱오차(integrated squared error)를 최소로 하는 것을 생각할 수 있다.

$$LMISE = \sum_{i=1}^{N} ||\varepsilon_i(t)||^2 = \sum_{i=1}^{N} \int [\mathrm{Prec}_i(t) - \alpha(t) - \int \mathrm{Temp}_i(s)\beta(s,t)ds]^2 dt \tag{3.3}$$

식 (3.3)의 기준에 맞는 추정을 위해서는 기저함수 표현을 이용하는 방법, 평활화 기법을 적용하는 방법 등 몇 가지가 있는데 본 연구에서는 기저함수 표현을 이용하였다. 간단한 설명을 위해, $x_i = \mathrm{Temp}_i$, $y_i = \mathrm{Prec}_i$라 하고, $x_i^* = x_i - \bar{x}$, $y_i^* = y_i - \bar{y}$라 하자. 그러면 식 (3.3)은

$$LMISE = \sum_{i=1}^{N} \int [y_i^*(t) - \int x_i^*(s)\beta(s,t)ds]^2 dt \tag{3.4}$$

로 나타낼 수 있으며, 이 경우에 $\alpha(t)$는 $\bar{y}(t) - \int x(s)\beta(s,t)ds$와 같다. 다음은 x_i^*와 y_i^*를 각각 기저 ϕ_j와 ψ_k로 표현하고, 회귀함수 β를 다음과 같이 나타낸다.

$$x_i^* = \sum_{j=1}^{J} c_{ij}\phi_j = c_i'\phi, \quad y_i^* = \sum_{k=1}^{K} d_{ik}\psi_k = d_i'\psi,$$
$$\beta(s,t) = \sum_{j=1}^{J}\sum_{k=1}^{K} b_{jk}\phi_j(s)\psi_k(t) = \phi(s)'\mathbf{B}\psi(t),$$

여기서 **B**는 계수 b_{jk}들을 포함하고 있는 $J \times K$ 행렬이고, J, K 값으로는 3.1 절에서 언급한대로 12를 사용하였다. 이를 이용하여 식 (3.4)를 행렬 표현으로 나타내고 특이값 분해 (Singular Value Decomposition, SVD)와 무어-펜로스의 일반화 역행렬 (Moore-Penrose generalized inverse)등의 행렬 이론을 이용하여 **B**를 구할 수 있다. 이에 관한 자세한 내용은 Ramsay와 Silverman (1997)의 11.2와 11.3절을 참고하기 바란다.

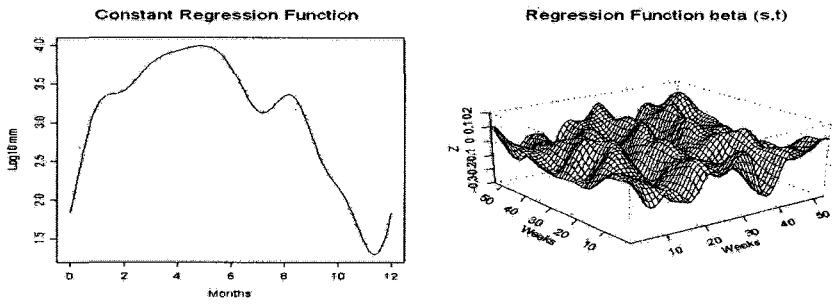

그림 3.13: 추정된 회귀계수 : 왼쪽은 $\hat{\alpha}(t)$이며 오른쪽은 $\hat{\beta}(s,t)$

그림 3.13은 우리나라 기온과 강수량 데이터를 이용해 함수적 선형모형 (3.2)에 적용한 결과인 절편 함수와 회귀계수 함수의 추정치를 나타낸다. 추정된 절편 함수 $\hat{\alpha}(t)$와 회귀계수 $\hat{\beta}(s,t)$ 함수를 이용하여 서울,철원, 충주, 제주지역의 자료에 각각 적용해 본 결과 강수량의 추정치는 그림 3.14와 같다. 추정치를 더욱 부드럽게 하고 싶으면 기저의 수를 줄이는 방법이나 스플라인 평활의 경우처럼 부드러움의 정도를 벌점화(penalized)하는 평활량을 도입하여 사용할 수 있다.

4. 결론

본 연구에서는 기온 데이터와 강수량 데이터를 퓨리에 기저에 기초한 함수적 데이터로 나타낸 후 평균, 분산, 상관관계, 주성분 분석 및 함수적 회귀분석을 실시해 보았다. 함수적 데이터 분석은 연속적인 성질에도 불구하고 이산형 형태로 관측된 관측치를 정확한 함수의 형태로 근사하여 분석하는 방법이라고 할 수 있다. 함수적 데이터 분석의 이론들은 전통적인 다변량 분석의 기본 아이디어를 확장하여 표현할 수 있다. 본 연구에서 진행한 함수적 분산분석의 결과에 따르면 일조권에 따라 우리나라를 6개 권역으로 구분한 것이 기온과 강수량에 의해서도 유의하다는 것을 확인하였다. 또한 기온에 따른 강수량의 변화를 추정하기 위한 회귀모형을 적합시켜 추정된 결과를 비교하였다. 본 연구에서 회귀모형의 유의성에 대한 언급은 기저로 표현할 때 의존하는 몇 개의 주관적인 모수 결정에 대한 의존 때문에 피하였다. 따라서, 본 연구에 이은 추후 과제로 함수적 데이터로 회귀모형을 적합시키

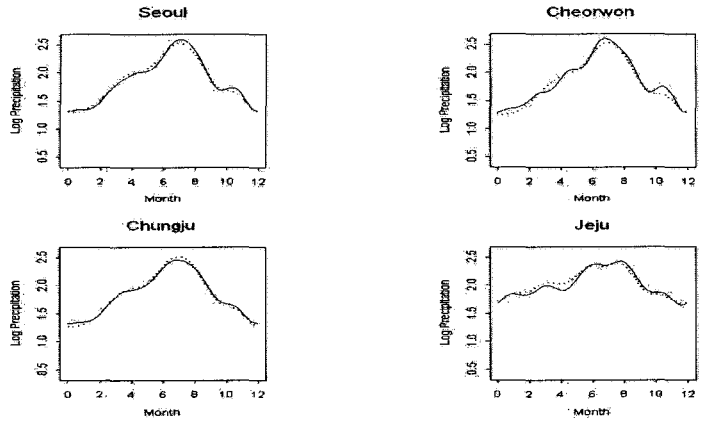

그림 3.14: 몇 개 지역에서 관측된 강수량 데이터(점선)와 선형모형에 의한 추정치(실선)

는 경우에 모형의 평가에 대한 연구와 반응변수와 설명변수 모두가 함수적 데이터로 표현된 회귀모형에서 데이터를 부드럽게 하는 평활량의 선택에 대한 연구가 필요하다고 본다.

참고문헌

문영수(1996). 일조시간의 연변화에 따른 한국의 지역 구분, <한국환경과학회지>, **5**, 253-263.

정창희(1976). 기후개요와 그 요인, *한국의 기후*, 김광식 외, 일지사, 10-12.

Locantore, N., Marron, J. S., Simpson, D. G., Tripoli, N., Zhang, J. T. and Cohen, K. L.(1999). Robust principal component analysis for functional data, *TEST*, **8**, 1-73.

Ramsay, J. O. and Dalzell, C. J.(1991). Some tools for functional data analysis, *Journal of the Royal Statistical Society*, Series B, **53**, 539-572.

Ramsay, J. O. and Silverman, B. W.(1997). *Functional Data Analysis,* Springer, New York.

Ramsay, J. O. and Silverman, B. W.(2002). *Applied Functional Data Analysis*, Springer, New York.

Ramsay, J. O. and Silverman, B. W.(2005). *Functional Data Analysis, Second Edition,* Springer, New York.

[2006년 5월 접수, 2006년 8월 채택]

Functional Data Analysis of Temperature and Precipitation Data*

Kee-Hoon Kang[1] Hong Se Ahn[2]

ABSTRACT

In this paper we review some methods for analyzing functional data and illustrate real application of functional data analysis. Representing methods for functional data by using basis function, analyzing functional variation by functional principal component analysis and functional linear models are reviewed. For a real application, we use temperature and precipitation data measured in Korea from the January of 1970 to the May of 2004. We apply functional principal component analysis for each data and test the significance of regional division done by using shining hours. We also estimate functional regression model for temperature and precipitation.

Keywords: functional principal component analysis, functional analysis of variance, functional linear models

* This work was supported by the SRC/ERC program of MOST/KOSEF (R11-2000-073-00000)

1) Associate Professor, Department of Statistics, Hankuk University of Foreign Studies, Yongin 449-791, Korea.
E-mail: khkang@hufs.ac.kr

2) Graduate Student, Department of Statistics, Hankuk University of Foreign Studies, Seoul 130-791, Korea.
E-mail: babaman@empal.com