The Korean Journal of Applied Statistics
2023, Vol. 36, No. 5, 473–488

DOI: https://doi.org/10.5351/KJAS.2023.36.5.473
Print ISSN1225-066X / Online ISSN 2383-5818

Nonparametric compositional data analysis for tourism industry in Gangwon area

Seongeun Park[a], Jeong Min Jeon[b], Young Kyung Lee[1,a]

[a]Department of Statistics, Kangwon National University;
[b]Department of Statistics, Seoul National University

Abstract

Gangwon-do is one of Korea's most popular tourist destinations, with varying tourism demands and trends across its subregions. It is crucial to identify the characteristics of tourism in each area and compare the tourism patterns over time to devise policies that revitalize tourism in each local government and promote balanced development across regions. In this paper, we classify the regions in Gangwon-do based on tourism data from the last four years and analyze the tourism pattern of each region using the non-Euclidean additive model proposed by Jeon *et al*. (2021). The model incorporates the proportions of visitors by age groups and the proportions of navigation searches by destination types as two covariates, and the proportions of tourism expenditure types as a response variable. We estimate the model using the smooth-backfitting method and coordinate-wise bandwidth selection. The results are visualized in ternary plots, and changes in tourism patterns over time are analyzed by comparing the ratios of prediction errors to fitting errors.

Keywords: compositional data, coordinate-wise bandwidth selection, non-Euclidean additive model, smooth backfitting, simplex

1. 서론

관광산업은 '굴뚝 없는 공장', '보이지 않는 무역'이라고 불리며 제품을 생산하는 공장이 없어도 고용 창출의 효과를 낼 수 있고, 외화 획득, 국제 친선, 문화교류 등의 역할을 하는 고부가가치 산업이다. 강원도는 우수한 자연자원을 보유한 국내 대표 관광지로 국가통계포털(KOSIS) 2020년 자료를 기준으로 지역 내 총생산(GRDP) 중 관광산업이 차지하는 비중이 16.3%, 서비스업이 차지하는 비중이 72.3%에 달하는 지역이다. 특히, 강원도는 2023년 6월부터 강원특별자치도로 변경됨에 따라 관광산업에 대한 기대가 더욱 커지고 있다.

COVID-19의 전 세계적인 유행은 사회, 문화, 보건, 경제, 기술 분야 등 지구촌 전반에 직간접적인 영향을 미쳤으며, 특히 타 산업과 밀접한 연관성을 가지고 있는 관광산업 분야에 4조 달러가 넘는 경제적 손실을 유발하였다 (UNCTAD, 2021). 이에 관광 분야의 다수 전문가 혹은 전문기관들은 COVID-19로 유발된 여행패턴 변화에 대한 분석이 관광산업의 조기 회복전략을 수립하는데 중요한 기초자료가 될 것이라 주장하며 국가 및 지역 수준에서 관광시장 회복을 위한 다양한 정책 및 기술전략 등을 모색하고 있다 (Badr 등 2020).

This Research was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No. NRF-2021R1A2C1003920).

[1]Corresponding author: Department of Statistics, Kangwon National University, 1 Gangwondaehakgil, Chuncheon-si, Gangwon-do 24341, Korea. E-mail: younklee@kangwon.ac.kr

Published 31 October 2023 / journal homepage: http://kjas.or.kr

한국관광공사 데이터랩(https://datalab.visitkorea.or.kr)에서 수집한 2018년부터 2021년까지의 자료에 따르면, 2019년 강원도를 방문한 방문자 수는 전년도 대비 4.48% 증가하였지만, 2020년 COVID-19로 인해 방문자 수가 5.6% 감소하였다. 가장 크게 방문자 수가 감소한 정선군은 27.43% 감소하였지만, 양양군은 오히려 10.29% 증가하였다. 또한, 2021년 기준 가장 많은 관광객이 방문한 강릉시는 3,043만 9,048명이 방문하여 강원도 관광객의 13.12%를 차지하고 있지만, 가장 적은 관광객이 방문한 양구군은 280만 4,386명이 방문하여 강원도 관광객의 1.21%만을 차지하고 있다. 이처럼 강원도 관광객은 강릉시 등 일부 지역에만 편중되어 있고, 연도에 따라 다른 방문자 추이를 보인다. 따라서 지역 간 균형적인 발전과 각 지자체가 관광 활성화 정책을 수립하는 데 있어 각 지역의 관광 특성을 파악하고, 연도별 관광 형태를 비교하는 것이 필요하다.

본 논문의 2장에서는 분석에 사용할 방법론을 소개한다. 3장에서는 한국관광공사에서 제공하는 강원도 각 행정지역별 관광객 자료들 중 관광지출액, 방문자 수, 내비게이션 검색 건수 자료를 소개한다. 4장에서는 강원도 방문객의 방문추이에 따라 지역군을 나누고, 각 지역군의 연도별 관광패턴을 분석한다. 특별히 COVID-19로 인한 관광패턴 변화에 초점을 맞춘다. 지역군을 나누기 위해 유클리디안(Euclidean) 데이터에 대한 군집분석을 활용하고, 지역군 및 연도별 관광패턴 파악을 위해 구성비(compositional) 데이터에 대한 비모수적 회귀모형을 적합시킨다.

2. 방법론

2.1. k-means 군집분석

군집분석은 관측치들을 몇 개의 군집으로 나누는 다변량 기법이다. k-means 군집분석은 비계층적 군집분석 방법 중 하나인데, 주어진 군집 수 k에 대해 각 군집은 하나의 중심을 가지고, 각 관측치는 가장 가까운 중심에 할당되어 같은 중심에 할당된 관측치들이 하나의 군집을 형성한다. k-means 군집분석은 계산량이 적어 대용량 데이터를 빠르게 처리할 수 있다는 장점이 있다.

2.2. 심플렉스

자연수 $D > 1$에 대해 다음과 같은 공간을 정의하자 (Egozcue 등, 2003).

$$S^D = \left\{ \boldsymbol{x} = [x_1, \ldots, x_D] \in \mathbb{R}^D \mid x_i > 0,\ \ i = 1, \ldots, D, \sum_{i=1}^{D} x_i = 1 \right\}.$$

이러한 공간을 심플렉스라고 부르고, 심플렉스의 원소를 구성비 벡터라고 부른다. 일부 성분이 0인 벡터는 심플렉스의 원소는 아니지만 이러한 데이터는 Pawlowsky-Glahn과 Buccianti (2011)에서 정리된 방법으로 분석할 수 있다. 심플렉스는 유클리디안 덧셈과 유클리디안 상수 곱에 대해 닫혀 있지 않기 때문에 심플렉스에서 값을 가지는 소위 구성비 데이터를 다루기 위해서는 특수한 연산자가 필요하다. 이를 위해 공간 $\mathbb{R}_+^D = \{\boldsymbol{z} = [z_1, \ldots, z_D] \in \mathbb{R}^D | z_i > 0,\ i = 1, \ldots, D\}$에 대해 작용소 $C : \mathbb{R}_+^D \to S^D$를 다음과 같이 정의하자.

$$C(\boldsymbol{z}) = \left[\frac{z_1}{\sum_{i=1}^{D} z_i}, \ldots, \frac{z_D}{\sum_{i=1}^{D} z_i} \right].$$

그리고 $\boldsymbol{x}, \boldsymbol{y} \in S^D$와 $\alpha \in \mathbb{R}$에 대해 다음과 같은 연산자와 내적을 정의하자 (Egozcue 등, 2003).

1) 벡터 덧셈

$$\boldsymbol{x} \oplus \boldsymbol{y} = C\left([x_1 y_1, \ldots, x_D y_D]\right).$$

2) 상수 곱

$$\alpha \odot \boldsymbol{x} = C\left([x_1^{\alpha}, \ldots, x_D^{\alpha}]\right).$$

3) 내적

$$< \boldsymbol{x}, \boldsymbol{y} > = \frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \log \frac{x_i}{x_j} \log \frac{y_i}{y_j}.$$

그러면 대응되는 영벡터(zero vector) $\mathbf{0}$와 벡터 뺄셈 연산자 $\ominus$는 다음과 같이 주어진다.

$$\mathbf{0} = [1/D, \ldots, 1/D], \quad \boldsymbol{x} \ominus \boldsymbol{y} = \boldsymbol{x} \oplus (-1 \odot \boldsymbol{y}).$$

그리고 크기(norm)와 거리(distance)는 다음과 같이 주어진다.

$$\|\boldsymbol{x}\| = \sqrt{\frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \left(\log \frac{x_i}{x_j}\right)^2}, \quad \|\boldsymbol{x} \ominus \boldsymbol{y}\| = \sqrt{\frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \left(\log \frac{x_i}{x_j} - \log \frac{y_i}{y_j}\right)^2}.$$

이와 같은 정의의 직관과 유용성은 Pawlowsky-Glahn과 Buccianti (2011)에서 보다 자세히 확인할 수 있다. 한편, S^D와 $\mathbb{R}^{D-1}$ 사이에는 isometric log-ratio 변환이 존재하는데 이를 정의하기 위해 다음과 같은 구성비 벡터를 정의하자.

$$\boldsymbol{e}_i = C\left(\left[\exp\left(\sqrt{\frac{1}{i(i+1)}}\right), \ldots, \exp\left(\sqrt{\frac{1}{i(i+1)}}\right), \exp\left(-\sqrt{\frac{i}{i+1}}\right), 1, \ldots, 1\right]\right), \quad i = 1, \ldots, D-1,$$

여기서, $\exp((i(i+1))^{-1/2})$은 i번 반복되고, 1은 $(D-i-1)$번 반복된다. 그러면 $\{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_{D-1}\}$은 S^D의 정규직교 기저(orthonormal basis)를 이룬다. 이제 isometric log-ratio 변환 $ilr_D : S^D \to \mathbb{R}^{D-1}$은 다음과 같이 정의된다 (Egozcue 등, 2003).

$$ilr_D(\boldsymbol{x}) = [< \boldsymbol{x}, \boldsymbol{e}_1 >, \ldots, < \boldsymbol{x}, \boldsymbol{e}_{D-1} >]. \tag{2.1}$$

이 변환은 centered log-ratio 변환, additive log-ratio 변환 등과 달리 S^D와 $\mathbb{R}^{D-1}$ 사이의 등거리 동형 사상(isometric isomorphism)이기 때문에 (Egozcue 등, 2003) 이를 이용해 다음 절에서 등장하는 심플렉스 위에서의 적분을 쉽게 정의하고 구현할 수 있다.

2.3. 비유클리디안 변수에 대한 가법모형

본 논문에서는 자료 분석을 위해 Jeon 등 (2021)이 제안한 비유클리디안(non-Euclidean) 변수에 대한 가법모형(additive model)을 이용한다. 특별히 다음과 같은 가법모형을 이용한다.

$$\boldsymbol{Y} = \mathrm{E}(\boldsymbol{Y}) \oplus \mathbf{f}_1(\boldsymbol{X}_1) \oplus \mathbf{f}_2(\boldsymbol{X}_2) \oplus \boldsymbol{\epsilon}, \tag{2.2}$$

여기서, $\boldsymbol{Y}$는 종속 변수로서 업종별 관광지출 비율을 나타내고, $\boldsymbol{X}_j$는 공변량으로서 $\boldsymbol{X}_1$은 연령대별 방문자 비율, $\boldsymbol{X}_2$는 유형별 검색 비율을 나타낸다. $\boldsymbol{Y}, \boldsymbol{X}_1, \boldsymbol{X}_2$가 값을 가지는 심플렉스를 각각 S^{D_Y}, S^{D_1}, S^{D_2}라고 하면, $\boldsymbol{\epsilon}$은 S^{D_Y}에서 값을 가지고 $\mathrm{E}(\boldsymbol{\epsilon}|\boldsymbol{X}) = \mathbf{0}$를 만족하는 오차항이고, $1 \le j \le 2$에 대해 $\mathbf{f}_j : S^{D_j} \to S^{D_Y}$는 $\mathrm{E}(\|\mathbf{f}_j(\boldsymbol{X}_j)\|^2) < \infty$를 만족하는 미지의 성분 함수(component function)이다. $\mathbf{c}_1, \mathbf{c}_2 \in S^{D_Y}$가 $\mathbf{c}_1 \oplus \mathbf{c}_2 = \mathbf{0}$를 만족하는 임의의 상수일 때, $\mathbf{f}_1(\boldsymbol{X}_1) \oplus \mathbf{f}_2(\boldsymbol{X}_2) = (\mathbf{f}_1(\boldsymbol{X}_1) \oplus \mathbf{c}_1) \oplus (\mathbf{f}_2(\boldsymbol{X}_2) \oplus \mathbf{c}_2)$이기 때문에 성분 함수의 식별가능성(identifiability)을 위해

조건 $\mathrm{E}(\mathbf{f}_1(\boldsymbol{X}_1)) = \mathrm{E}(\mathbf{f}_2(\boldsymbol{X}_2)) = \mathbf{0}$을 두자. 이러한 조건이 $\mathbf{f}_j$가 영함수(zero function)임을 암시하지는 않기 때문에 이는 $\boldsymbol{X}_j$가 $\boldsymbol{Y}$에 설명력이 없음을 나타내지는 않는다.

성분 함수 $\mathbf{f}_j$의 추정을 위해 모형 (2.2)의 양변에 $\mathrm{E}(\cdot|\boldsymbol{X}_j = \boldsymbol{x}_j), 1 \le j \le 2$를 씌워 정리하면 다음과 같은 연립 적분 방정식을 얻는다.

$$\begin{aligned}\mathbf{f}_1(\boldsymbol{x}_1) &= \mathbf{m}_1(\boldsymbol{x}_1) \ominus \mathrm{E}(\boldsymbol{Y}) \ominus \int_{S^{D_2}} \frac{p_{12}(\boldsymbol{x}_1, \boldsymbol{x}_2)}{p_1(\boldsymbol{x}_1)} \odot \mathbf{f}_2(\boldsymbol{x}_2)\, d\mu_2(\boldsymbol{x}_2), \\ \mathbf{f}_2(\boldsymbol{x}_2) &= \mathbf{m}_2(\boldsymbol{x}_2) \ominus \mathrm{E}(\boldsymbol{Y}) \ominus \int_{S^{D_1}} \frac{p_{12}(\boldsymbol{x}_1, \boldsymbol{x}_2)}{p_2(\boldsymbol{x}_2)} \odot \mathbf{f}_1(\boldsymbol{x}_1)\, d\mu_1(\boldsymbol{x}_1),\end{aligned} \tag{2.3}$$

여기서, $\mathbf{m}_j(\boldsymbol{x}_j) = \mathrm{E}(\boldsymbol{Y}|\boldsymbol{X}_j = \boldsymbol{x}_j)$이고, 적분은 S^{D_Y}에서 값을 가지는 함수에 대한 보크너(Bochner) 적분이고 (Jeon과 Park, 2020), μ_j는 보렐(Borel) 집합 $A \subset S^{D_j}$과 르벡 측도(Lebesgue measure) Leb에 대해 $\mu_j(A) = \mathrm{Leb}(ilr_{D_j}(A))$로 정의되는 보렐 측도이고, p_{12}는 $(\boldsymbol{X}_1, \boldsymbol{X}_2)$의 $\mu_1 \otimes \mu_2$에 대한 결합확률밀도함수이고, p_j는 $\boldsymbol{X}_j$의 주변확률밀도함수이다. 여기서, $\mu_1 \otimes \mu_2$는 μ_1와 μ_2의 곱 측도(product measure)를 나타낸다. 본 논문에서는 식 (2.3)의 $\mathrm{E}(\boldsymbol{Y})$와 함수 p_{12}, p_j, $\mathbf{m}_j$를 추정한 후, 이들을 식 (2.3)에 대입하여 얻어지는 연립 적분 방정식의 해를 찾는 평활역적합(smooth backfitting) 방법을 이용한다 (Mammen 등, 1999; Jeon과 Park, 2020; Jeon 등, 2021). 이를 위해 다음과 같은 정규화 커널(normalized kernel)을 정의하자.

$$K_{h_j}\left(\boldsymbol{x}_j, \boldsymbol{u}_j\right) = \frac{K\left(\left\|\boldsymbol{x}_j \ominus \boldsymbol{u}_j\right\| / h_j\right)}{\int_{S^{D_j}} K\left(\left\|\boldsymbol{v}_j \ominus \boldsymbol{u}_j\right\| / h_j\right) d\mu_j(\boldsymbol{v}_j)},$$

여기서, h_j는 띠폭(bandwidth)이고, K는 $[0, 1)$에서 0보다 크고 $[1, \infty)$에서 0인 임의의 연속 함수이다. 본 논문에서는 Epanechnikov 커널을 K로 사용한다. 한편, S^{D_j} 위에서 μ_j에 대한 적분은 Jeon 등 (2021)의 3.3.1.2절을 바탕으로 구현할 수 있다. 이제 n개의 관측치 $\{(\boldsymbol{Y}^i, \boldsymbol{X}_1^i, \boldsymbol{X}_2^i) : 1 \le i \le n\}$가 주어졌을 때, $\mathrm{E}(\boldsymbol{Y})$, p_{12}, p_j, $\mathbf{m}_j$를 다음과 같이 추정한다.

$$\begin{aligned}\bar{\boldsymbol{Y}} &= \frac{1}{n} \odot \bigoplus_{i=1}^{n} \boldsymbol{Y}^i, \\ \hat{p}_{12}(\boldsymbol{x}_1, \boldsymbol{x}_2) &= \frac{1}{n} \sum_{i=1}^{n} K_{h_1}\left(\boldsymbol{x}_1, \boldsymbol{X}_1^i\right) K_{h_2}\left(\boldsymbol{x}_2, \boldsymbol{X}_2^i\right), \\ \hat{p}_j(\boldsymbol{x}_j) &= \frac{1}{n} \sum_{i=1}^{n} K_{h_j}\left(\boldsymbol{x}_j, \boldsymbol{X}_j^i\right), \\ \hat{\mathbf{m}}_j(\boldsymbol{x}_j) &= \left(n\, \hat{p}_j(\boldsymbol{x}_j)\right)^{-1} \odot \bigoplus_{i=1}^{n} K_{h_j}\left(\boldsymbol{x}_j, \boldsymbol{X}_j^i\right) \odot \boldsymbol{Y}^i.\end{aligned}$$

그러면 성분 함수 $\mathbf{f}_j$의 평활역적합 추정량은 연립 적분 방정식

$$\begin{aligned}\hat{\mathbf{f}}_1(\boldsymbol{x}_1) &= \hat{\mathbf{m}}_1(\boldsymbol{x}_1) \ominus \bar{\boldsymbol{Y}} \ominus \int_{S^{D_2}} \frac{\hat{p}_{12}(\boldsymbol{x}_1, \boldsymbol{x}_2)}{\hat{p}_1(\boldsymbol{x}_1)} \odot \hat{\mathbf{f}}_2(\boldsymbol{x}_2)\, d\mu_2(\boldsymbol{x}_2), \\ \hat{\mathbf{f}}_2(\boldsymbol{x}_2) &= \hat{\mathbf{m}}_2(\boldsymbol{x}_2) \ominus \bar{\boldsymbol{Y}} \ominus \int_{S^{D_1}} \frac{\hat{p}_{12}(\boldsymbol{x}_1, \boldsymbol{x}_2)}{\hat{p}_2(\boldsymbol{x}_2)} \odot \hat{\mathbf{f}}_1(\boldsymbol{x}_1)\, d\mu_1(\boldsymbol{x}_1)\end{aligned} \tag{2.4}$$

을 만족하고 $\int_{S^{D_j}} \hat{p}_j(\boldsymbol{x}_j) \odot \hat{\mathbf{f}}_j(\boldsymbol{x}_j)\, d\mu_j(\boldsymbol{x}_j) = \mathbf{0}$와 $\int_{S^{D_j}} \|\hat{\mathbf{f}}_j(\boldsymbol{x}_j)\|^2 \hat{p}_j(\boldsymbol{x}_j) d\mu_j(\boldsymbol{x}_j) < \infty$를 만족하는 함수 $\hat{\mathbf{f}}_j$로 주어진다. 그리고 $\mathrm{E}(\boldsymbol{Y}|\boldsymbol{X} = \boldsymbol{x})$의 추정량 $\hat{\mathrm{E}}(\boldsymbol{Y}|\boldsymbol{X} = \boldsymbol{x})$은 $\bar{\boldsymbol{Y}} \oplus \hat{\mathbf{f}}_1(\boldsymbol{x}_1) \oplus \hat{\mathbf{f}}_2(\boldsymbol{x}_2)$로 주어진다.

한편, $\hat{\mathbf{f}}_j$는 닫힌 형태로 주어지지 않기 때문에 이를 반복 알고리즘으로 찾아야 하는데 이를 위해 초기 함수 $\hat{\mathbf{f}}_j^{(0)}$를 다음과 같이 선택한다.

$$\hat{\mathbf{f}}_j^{(0)}(\boldsymbol{x}_j) = \frac{1}{n} \odot \bigoplus_{i=1}^{n} \left(\frac{K_{h_j}\left(\boldsymbol{x}_j, \boldsymbol{X}_j^i\right)}{\hat{p}_j(\boldsymbol{x}_j)} - 1 \right) \odot \boldsymbol{Y}^i.$$

그리고 $r = 1, 2, \ldots$에 대해 다음과 같이 이전 단계의 함수를 업데이트한다.

$$\hat{\mathbf{f}}_1^{(r)}(\boldsymbol{x}_1) = \hat{\mathbf{m}}_1(\boldsymbol{x}_1) \ominus \bar{\boldsymbol{Y}} \ominus \int_{S^{D_2}} \frac{\hat{p}_{12}(\boldsymbol{x}_1, \boldsymbol{x}_2)}{\hat{p}_1(\boldsymbol{x}_1)} \odot \hat{\mathbf{f}}_2^{(r-1)}(\boldsymbol{x}_2)\, d\mu_2(\boldsymbol{x}_2),$$
$$\hat{\mathbf{f}}_2^{(r)}(\boldsymbol{x}_2) = \hat{\mathbf{m}}_2(\boldsymbol{x}_2) \ominus \bar{\boldsymbol{Y}} \ominus \int_{S^{D_1}} \frac{\hat{p}_{12}(\boldsymbol{x}_1, \boldsymbol{x}_2)}{\hat{p}_2(\boldsymbol{x}_2)} \odot \hat{\mathbf{f}}_1^{(r)}(\boldsymbol{x}_1)\, d\mu_1(\boldsymbol{x}_1).$$

위 과정은 $\max_{1\le j\le 2} \int_{S^{D_j}} \|\hat{\mathbf{f}}_j^{(r)}(\boldsymbol{x}_j) \ominus \hat{\mathbf{f}}_j^{(r-1)}(\boldsymbol{x}_j)\|^2 \hat{p}_j(\boldsymbol{x}_j) d\mu_j(\boldsymbol{x}_j) < 10^{-4}$를 만족할 때까지 반복한다. Jeon 등 (2021)은 실제로 $\hat{\mathbf{f}}_j^{(r)}$이 r이 커짐에 따라 $\hat{\mathbf{f}}_j$로 수렴하고, $\hat{\mathbf{f}}_j$가 n이 커짐에 따라 $\mathbf{f}_j$로 수렴함을 증명한바 있다.

2.4. 성분별 띠폭 선택법(coordinate-wise bandwidth selection; CBS)

본 논문에서는 Jeon과 Park (2020)이 제안한 CBS 방법으로 띠폭 h_j를 선택한다. CBS 방법을 설명하기 위해 주어진 띠폭 h_1, h_2에 대해 5-fold 교차검증(cross-validation)값 $\mathrm{CV}(h_1, h_2)$을 다음과 같이 정의하자.

$$\mathrm{CV}(h_1, h_2) = \sum_{k=1}^{5} \sum_{i \in S_k} \left\| \boldsymbol{Y}^i \ominus \hat{\mathbf{m}}^{(-S_k)}\left(\boldsymbol{X}^i; h_1, h_2\right) \right\|^2,$$

여기서, $S_1, \ldots, S_5$는 $\cup_{k=1}^{5} S_k = \{1, \ldots, n\}$을 만족하는 서로 배반인 집합이고, $\hat{\mathbf{m}}^{(-S_k)}(\boldsymbol{x}; h_1, h_2)$는 S_k에 속하지 않는 관측치와 h_1, h_2를 이용해 얻은 $\mathrm{E}(\boldsymbol{Y}|\boldsymbol{X} = \boldsymbol{x})$의 추정량이다. 이제 띠폭 그리드(grid) $\prod_{j=1}^{2} \{g_{j1}, \ldots, g_{jL_j}\}$를 잡고, $\{g_{j1}, \ldots, g_{jL_j}\}$에서 임의의 초기값 $h_j^{(0)}$을 선택한다. 그리고 $t = 1, 2, \ldots$에 대해 다음의 과정을 $(h_1^{(t)}, h_2^{(t)}) = (h_1^{(t-1)}, h_2^{(t-1)})$을 만족할 때까지 반복한다.

$$h_1^{(t)} = \underset{g_1 \in \{g_{11}, \ldots, g_{1L_1}\}}{\arg\min}\ \mathrm{CV}\left(g_1, h_2^{(t-1)}\right),$$
$$h_2^{(t)} = \underset{g_2 \in \{g_{21}, \ldots, g_{2L_2}\}}{\arg\min}\ \mathrm{CV}\left(h_1^{(t)}, g_2\right).$$

3. 분석 자료

3.1. 자료 설명

본 논문에서는 한국관광공사 데이터랩(https://datalab.visitkorea.or.kr)의 2018년 1월 1일부터 2021년 12월 31일까지의 강원도 방문객 자료를 분석하였다. 자료는 크게 관광지출액, 방문자 수, 검색 건수, 이렇게 3가지로 구분되어 있다.

첫 번째로, 관광지출액 자료는 BC카드와 신한카드에서 제공된 것으로 현지인과 외지인을 포함하는 내국인을 대상으로 수집된 자료이다. 여기서, 현지인은 카드 발급자 자택 주소지가 분석 대상 지자체인 사람, 외지인은 그 외의 사람을 의미한다. 관광지출액 자료는 일별 데이터로 업종 분류기준에 따라 '쇼핑업', '숙박업', '식음료업', '여가서비스업', '여행업', '운송업' 총 6개의 업종에 따라 수집되었다. 본 분석에서는 외지인

Table 1: Variables used in data analysis

Variable	Variable description	Component	Category
$\boldsymbol{Y}^o$	Tourism expenditure	Y_1^o	food & beverage
		Y_2^o	accommodation
		Y_3^o	shopping
		Y_4^o	others
$\boldsymbol{X}_1^o$	Number of visitors	X_{11}^o	20s
		X_{12}^o	30s
		X_{13}^o	50s and order
		X_{14}^o	others
$\boldsymbol{X}_2^o$	Search volume	X_{21}^o	food
		X_{22}^o	accommodation
		X_{23}^o	shopping
		X_{24}^o	others

자료를 다루었고, 6개 업종들을 외지인의 관광특성을 잘 파악할 수 있는 지출 부문인 '쇼핑업', '숙박업', '식음료업'과 그 외 업종인 '기타 업종', 총 4개의 업종($\boldsymbol{Y}^o$)으로 다시 구분하였다.

두 번째로, 방문자 수 자료는 KT 이동통신에서 제공된 것으로 이 중 연령대 파악이 가능한 외지인 방문자 수를 대상으로 한다. 외지인 방문자란 기초지자체 기준 해당 지자체에 상주(거주/통학/통근 등)하고 있지 않은 사람으로 전월 기준 월 3회 이하 방문자를 의미한다. 방문자 수 자료는 월별 데이터로 '10대 미만', '10대', '20대', '30대', '40대', '50대', '60대', '70대 이상' 총 8개 연령대에 따라 수집됐다. 많은 지자체들이 관광활성화 정책의 주체로 관심을 갖는 연령대는 젊은 층으로 대표되는 MZ세대와 시니어세대이다. 본 분석에서는 이를 반영하여 연령대를 '20대', '30대', '50대 이상'과 그 외 연령대인 '기타 연령대', 총 4개의 연령대($\boldsymbol{X}_1^o$)로 다시 구분하였다.

세 번째로, 검색 건수 자료는 Tmap 모빌리티에서 제공된 자료로 Tmap 어플리케이션 사용자가 목적지를 조회하고 100m & 1분 이상 이동한 행위에 대한 건수를 나타낸다. 검색 건수 자료는 일별 데이터로 목적지 유형에 따라 '레저스포츠', '문화관광', '쇼핑', '숙박', '역사관광', '음식', '자연관광', '체험관광', '기타관광'으로 수집되었으며 이들을 $\boldsymbol{Y}^o$처럼 '쇼핑', '숙박', '음식'과 그 외 유형인 '기타 유형', 총 4개의 유형($\boldsymbol{X}_2^o$)으로 다시 구분하였다. 이러한 구분들은 Table 1에 정리되어 있다.

3.2. 자료 전처리

수집한 자료는 값 자체가 아닌 비율을 이용한 분석을 위해 구성비 벡터로 변환하였다. 관광지출액 자료는 구성비 벡터 $\boldsymbol{Y} = (Y_1, Y_2, Y_3, Y_4)$, $Y_\ell = Y_\ell^o / \sum_{k=1}^4 Y_k^o$, $\ell = 1, \ldots, 4$로 변환하였는데, 여기서 Y_1, Y_2, Y_3, Y_4는 각각 매일 전체 관광지출액 중 '식음료업', '숙박업', '쇼핑업', '기타 업종'이 차지하는 비율을 나타낸다. 마찬가지로 방문자 수 자료는 구성비 벡터 $\boldsymbol{X}_1 = (X_{11}, X_{12}, X_{13}, X_{14})$, $X_{1\ell} = X_{1\ell}^o / \sum_{\ell=1}^4 X_{1\ell}^o$, $\ell = 1, \ldots, 4$로 변환하였는데, $X_{11}, X_{12}, X_{13}, X_{14}$는 각각 매월 전체 방문자 중 '20대', '30대', '50대 이상', '기타 연령대'가 차지하는 비율을 나타낸다. 또한, 검색 건수 자료는 $\boldsymbol{X}_2 = (X_{21}, X_{22}, X_{23}, X_{24})$, $X_{2\ell} = X_{2\ell}^o / \sum_{\ell=1}^4 X_{2\ell}^o$, $\ell = 1, \ldots, 4$로 변환하였는데, $X_{21}, X_{22}, X_{23}, X_{24}$는 각각 매일 전체 검색 건수 중 '음식', '숙박', '쇼핑', '기타 유형'이 차지하는 비율을 나타낸다.

방문자 수 자료는 일별 데이터가 아닌 월별 데이터로 주어지기 때문에 모형 (2.2) 적합 시, 이를 각 월의 날짜 갯수 만큼 반복하여 일별 데이터로 변환하였다. 또한, 모형 (2.2)는 각 성분의 값이 0보다 큰 구성비 벡터만을 다룰 수 있기 때문에 강원도 18개 지역 중 쇼핑업 관광지출액 비율이 0인 양구군은 분석에서 제외하였다.

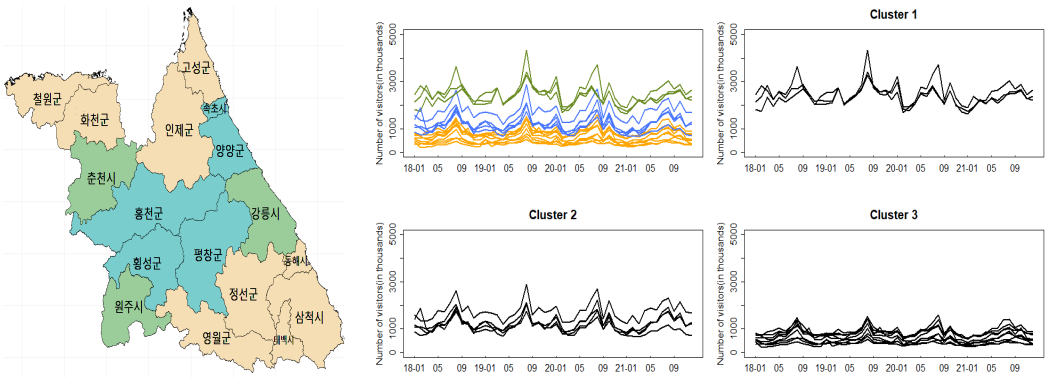

Figure 1: *Clustering results based on the number of visitors.*

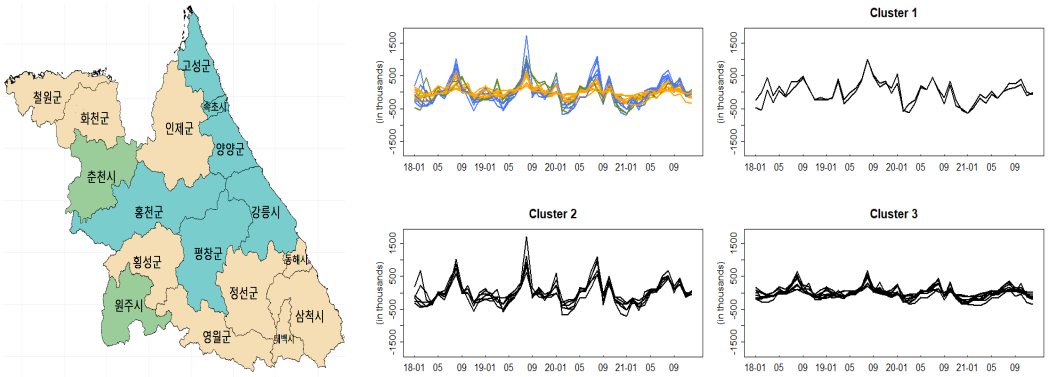

Figure 2: *Clustering results based on the visiting pattern.*

4. 분석 결과

4.1. 군집분석 결과

양구군을 제외한 17개 시/군 지역을 유클리디안 변수인 월평균 방문자 수와 월평균 방문자 증감 수에 따라 각각 군집화하였다. 이때, $k = 3$인 k-means 군집분석을 적용하였다.

1) 방문자 수에 따른 군집분석

Figure 1은 방문자 수에 따른 군집분석 결과를 나타낸다. 군집 1은 3개, 군집 2는 5개, 군집 3은 9개의 지역이 각각 속해있다. 각 그래프의 가로축은 날짜, 세로축은 방문자 수를 나타낸다.

군집 1은 방문자 수가 가장 많은 지역들로 구성된 군집이다. 특히 여름철에 방문자 수가 집중되어 있으며, 코로나19가 발생한 2020년부터 방문자 수가 감소하기 시작하여 코로나19가 지속된 2021년에는 여름철 방문자 수가 크게 감소하였다. 군집 2는 군집 1 다음으로 방문자 수가 많은 지역들로 이루어진 군집이다. 군집 1과 마찬가지로 여름철에 방문자 수가 증가하기 시작하여 매년 8월에 가장 많은 방문자 수를 보이고 9월에 다시 감소하는 형태를 보인다. 또한, 군집 2도 코로나19 이후에 방문자 수가 감소하는 경향을 보였다. 군집 3은 방문자 수가 가장 적은 지역들로 이루어진 군집으로 다른 군집들과 마찬가지로 여름철에 방문자

Table 2: ASFE values

	2018	2019	2020	2021
Cluster 1	0.2710	0.1785	0.2591	0.3538
Cluster 2	1.1059	0.7485	1.1242	1.1456
Cluster 3	1.7289	1.6542	1.7677	1.6946

Table 3: ASPE and ASPE/ASFE values

	Test dataset	
	2020	2021
Cluster 1	0.3734 (2.09)	0.4797 (2.69)
Cluster 2	0.9098 (1.22)	0.8633 (1.15)
Cluster 3	1.9234 (1.16)	2.0469 (1.24)

수가 증가하였지만, 그 증가폭이 상대적으로 작은 편이다. 또한, 코로나19 이후 방문자 수가 감소하는 형태를 보이나, 다른 군집들과 비교해 감소폭이 눈에 띄게 드러나지는 않는다. 군집 1~3 모두 각 군집 내에서 방문자 수가 많은 시점과 적은 시점이 지역별로 약간 차이가 있고, 고점과 저점의 높이도 지역별로 상당히 다른 것을 볼 수 있다.

2) 방문자 증감 수에 따른 군집분석

Figure 2의 오른쪽 그림들은 각 지역별로 월평균 방문자 수 대비 월별 방문자 증감 수를 나타낸다. 즉, 가로축은 날짜를 의미하고, 세로축은 (월별 방문자 수-월평균 방문자 수)를 의미한다. 그리고 이를 기준으로 얻은 3개의 군집이 Figure 2의 왼쪽 그림에 표현되어 있다. 군집 1은 2개, 군집 2는 6개, 군집 3은 9개의 도시가 각각 속해있다.

군집 1은 증가 · 감소를 자주 반복하고, 그 증감폭이 크지 않아 계절의 효과가 크지 않음을 알 수 있다. 그리고 이러한 증감폭은 코로나19 이후 더 줄어든 것을 알 수 있다. 군집 2는 변동이 가장 큰 지역으로 여름철과 겨울철에 방문자 수가 증가하는 특징이 있으며 특히 8월에 월평균 방문자 수 대비 큰 증가를 보인다. 코로나19 이후에도 계절에 따른 방문자 수 증감 패턴은 유지되지만 그 증감폭은 줄어든 것을 알 수 있다. 군집 3은 군집 2와 마찬가지로 매년 8월에 방문자 수가 집중되었다. 그러나 다른 군집에 비해 전반적으로 방문자 수 변동 폭이 작은데 이러한 특징은 코로나19 이후에 더 심화되었다.

위 두 가지 군집분석 결과를 비교해 보았을 때, 방문자 수에 기반한 첫 번째 군집분석 결과보다는 월별 방문자 수 증감에 기반한 두 번째 군집분석 결과가 각 군집내에서 지역간 차이가 적고, 각 군집을 구성하는 지역들의 특징을 해석하기 용이하였다. 따라서, 두 번째 군집분석에서 얻은 3개의 군집에 대해 각 군집별로 회귀분석을 진행하였다. 이 군집들은 구성하는 지역들의 특징을 고려하여 행정도시 지역군(군집 1), 유명관광도시 지역군(군집 2), 그외 소도시 지역군(군집 3)으로 생각할 수 있다.

4.2. 회귀분석 결과(방문자 수 증감에 따른 군집)

각 모형에 대해 평균 제곱 적합 오차(average squared fitting error; ASFE)와 평균 제곱 예측 오차(average squared prediction error; ASPE)를 구하여 비교하였다. ASFE는 훈련자료(train data)에서 적합된 업종별 관광지출액 비율($\hat{\boldsymbol{Y}}$)과 훈련자료 내의 실제 업종별 관광지출액 비율($\boldsymbol{Y}$)의 차이에 대한 제곱의 평균으로 정의된다. 즉, I_{tr}을 훈련자료의 인덱스(index) 집합이라고 하고, $|I_{\text{tr}}|$을 I_{tr}의 원소 수라고 할 때, ASFE는 아래와 같이

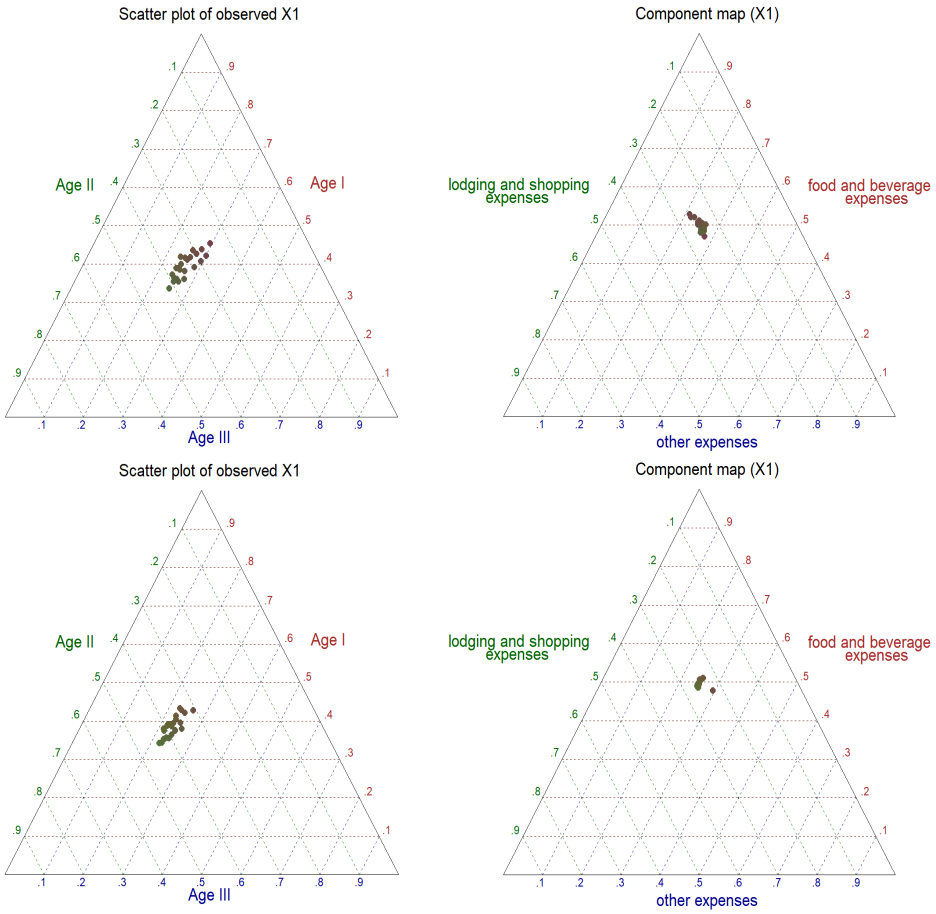

Figure 3: *Ternary plots of* $\boldsymbol{X}_1^a$*(left) and* $\hat{\mathbf{f}}_1(\boldsymbol{X}_1^a)$*(right) for the years 2019(top) and 2021(bottom) in the* 1^{st} *cluster.*

계산된다.

$$\text{ASFE} = \frac{1}{|I_{\text{tr}}|}\sum_{i\in I_{\text{tr}}}\left\|\boldsymbol{Y}^i \ominus \hat{\boldsymbol{Y}}^i\right\|^2 = \frac{1}{|I_{\text{tr}}|}\sum_{i\in I_{\text{tr}}}\frac{1}{2\cdot 4}\sum_{k=1}^{4}\sum_{l=1}^{4}\left(\log\frac{Y_k^i}{Y_l^i} - \log\frac{\hat{Y}_k^i}{\hat{Y}_l^i}\right)^2.$$

ASPE는 훈련자료로 추정한 회귀 함수를 이용해 시험자료(test data) 내의 업종별 관광지출액 비율을 예측한 값과 시험자료 내의 실제 업종별 관광지출액 비율의 차이에 대한 제곱의 평균으로 정의된다. 즉, I_{te}를 $\{1,\ldots,n\}\setminus I_{\text{tr}}$라고 하고, $|I_{\text{te}}|$를 I_{te}의 원소 수라고 할 때, ASPE는 아래와 같이 계산된다.

$$\text{ASPE} = \frac{1}{|I_{\text{te}}|}\sum_{i\in I_{\text{te}}}\left\|\boldsymbol{Y}^i \ominus \hat{\boldsymbol{Y}}^i\right\|^2.$$

ASFE 값을 통해 회귀모형이 잘 적합되었는지 파악하고, ASPE 값을 통해 연도별 관광패턴에 대한 예측의 정확도를 파악하고자 한다.

Table 2는 각 군집에 대한 연도별 ASFE 값을 나타낸 것이다. 모든 군집에서 2019년도 자료의 ASFE 값이 가장 작기 때문에, 가정한 회귀모형이 2019년도 자료에서 가장 잘 적합된다고 볼 수 있다. 따라서, 각 군집에

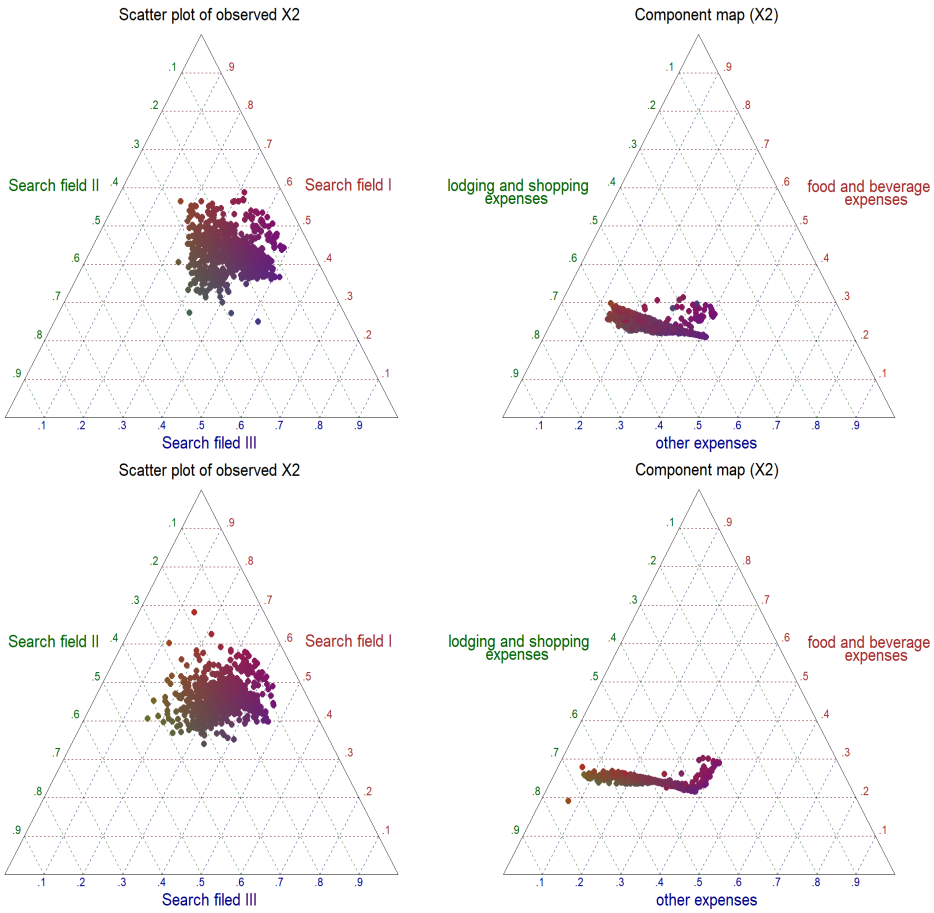

Figure 4: *Ternary plots of* $\boldsymbol{X}_2^a$ *(left) and* $\hat{\boldsymbol{t}}_2(\boldsymbol{X}_2^a)$ *(right) for the years 2019 (top) and 2021 (bottom) in the* 1^{st} *cluster.*

대해 2019년도 자료를 훈련자료로 잡고, 이후 연도인 2020년도와 2021년도의 자료를 시험자료로 한 ASPE 값들을 구해보았다.

Table 3의 괄호 밖은 ASPE 값을 나타내고, 괄호 안은 ASFE 대비 ASPE의 비율인 ASPE/ASFE를 나타낸다. 서로 다른 군집들의 ASPE 값을 비교하면, 군집 1의 값이 제일 작기 때문에 예측의 정확도 측면에서 군집 1의 정확도가 제일 높다고 볼 수 있으나, 이는 각 군집을 구성하는 지역의 수가 다르고, 따라서 표본의 수가 다르기 때문에 정확한 해석으로 볼 수 없다. 따라서 ASPE 값 자체보다 비율 ASPE/ASFE 값을 살펴보면, 군집 1의 비율이 다른 군집들의 비율보다 훨씬 크기 때문에 코로나 19로 인해 군집 1의 관광패턴이 가장 크게 변하였다고 볼 수 있다. 이제 COVID-19 직전인 2019년도 자료와 관광패턴이 크게 변한 2021년도 자료에 대해 각각 모형 추정 후 성분함수 추정량을 비교함으로서 COVID-19의 영향을 구체적으로 살펴보고자 한다.

성분함수 추정량이 S^3에서 값을 가지는 경우 이를 삼각 도표(ternary plot)에 표현하여 성분함수 추정량을 쉽게 비교할 수 있다. 따라서 S^4에서 값을 가지는 기존의 비율 자료를 S^3에서 값을 가지는 비율 자료로 변환하였다. 이를 위해, 기존의 관광지출액 비율 자료 $\boldsymbol{Y} = (Y_1, Y_2, Y_3, Y_4)$의 범주를 조정하여 '식음료업', '숙박업 및 쇼핑업', '기타업종'이 차지하는 비율 자료인 $\boldsymbol{Y}^a = (Y_1, (Y_2 + Y_3), Y_4)$를 얻었다. 또한, 결과 해석의 편의를 위해 $\boldsymbol{X}_1$과 $\boldsymbol{X}_2$도 S^3에서 값을 가지는 비율 자료로 변환하였는데, 연령대별 방문자 수 비율 자료 $\boldsymbol{X}_1$ =

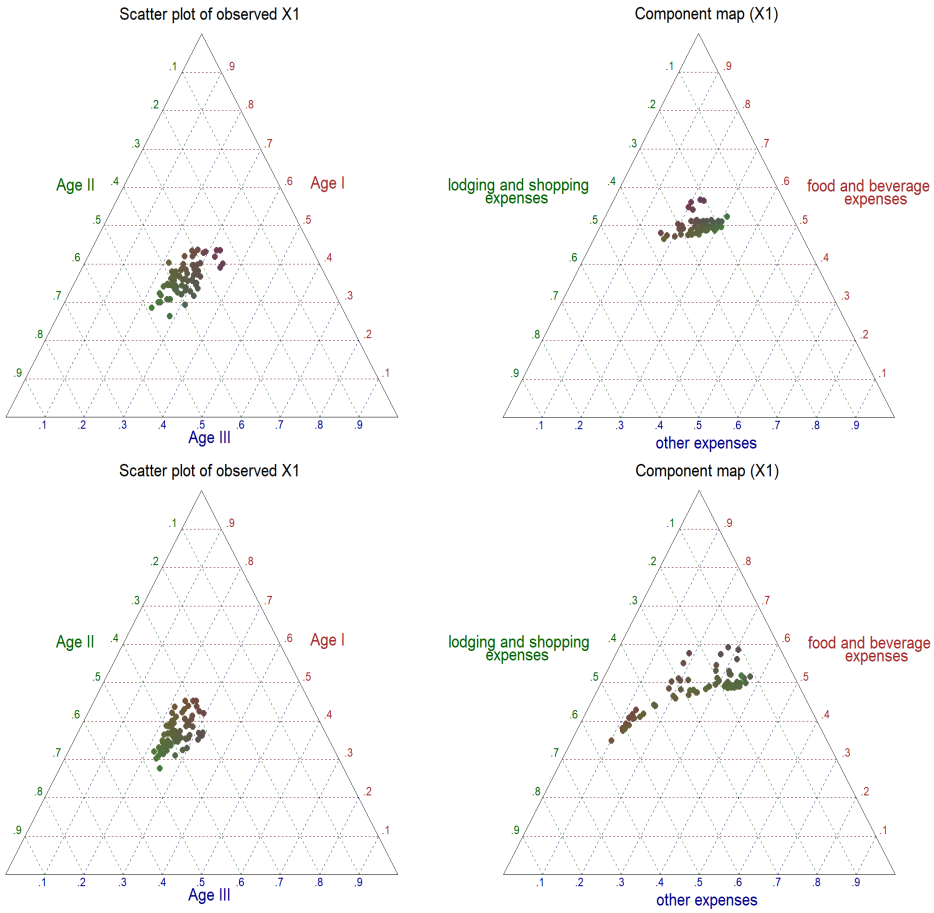

Figure 5: *Ternary plots of* $\boldsymbol{X}_1^a$ *(left) and* $\hat{\mathbf{f}}_1(\boldsymbol{X}_1^a)$ *(right) for the years 2019 (top) and 2021 (bottom) in the* 2^{nd} *cluster.*

$(X_{11}, X_{12}, X_{13}, X_{14})$는 $\boldsymbol{X}_1^a = ((X_{11} + X_{12}), X_{13}, X_{14})$로 변환하였다. 여기서, $\boldsymbol{X}_1^a$는 'age I(20-30대)', 'age II(50대 이상)', 'age III(기타 연령대)'가 차지하는 비율 자료를 나타낸다. 그리고 분야별 검색 건수 비율 자료 $\boldsymbol{X}_2 = (X_{21}, X_{22}, X_{23}, X_{24})$는 $\boldsymbol{X}_2^a = (X_{21}, (X_{22} + X_{23}), X_{24})$로 변환하였다. 여기서, $\boldsymbol{X}_2^a$는 'search field I(음식)', 'search field II(숙박과 쇼핑)', 'search field III(기타 유형)'이 차지하는 비율 자료를 나타낸다. 그리고 모형 (2.2)에서 $(\boldsymbol{Y}, \boldsymbol{X}_1, \boldsymbol{X}_2)$ 대신 $(\boldsymbol{Y}^a, \boldsymbol{X}_1^a, \boldsymbol{X}_2^a)$를 대입한 모형을 추정하였다. Figures 3–8에서 왼쪽 삼각 도표는 $\boldsymbol{X}_j^a$의 산점도이고, 오른쪽 삼각 도표의 점들은 $\hat{\mathbf{f}}_j(\boldsymbol{X}_j^a)$의 값들이다. $\hat{\mathbf{f}}_j(\boldsymbol{X}_j^a)$ 값의 색상은 왼쪽 삼각 도표에 있는 $\boldsymbol{X}_j^a$의 색상으로 주어진다. 삼각 도표 상의 각 점은 S^3에 속하는 구성비 벡터인데 벡터의 첫 번째 값은 빨간색 점선을 따라 오른쪽 수평으로 이동해서 닿는 지점의 값이고, 두 번째 값은 초록색 점선을 따라 왼쪽 위 방향으로 이동해서 닿는 지점의 값이며, 세 번째 값은 파란색 점선을 따라 왼쪽 아래 방향으로 이동해서 닿는 지점의 값이다. 세 값의 합은 언제나 1이 된다. 왼쪽 삼각 도표 그림에서 점의 각 색상이 선명할수록 색상에 해당하는 범주의 비율이 다른 범주들의 비율보다 월등히 큰 것을 나타낸다. 예를 들어, 점의 색상이 빨강과 초록 사이의 색상인 경우는 오른쪽과 왼쪽 범주의 비율이 비슷함을 나타낸다. 오른쪽 삼각 도표 그림의 점의 색상은 왼쪽 그림의 구성비 자료의 색상을 그대로 유지하여, 추정된 성분함수 값 위치로 이동을 확인할 수 있게 나타냈다.

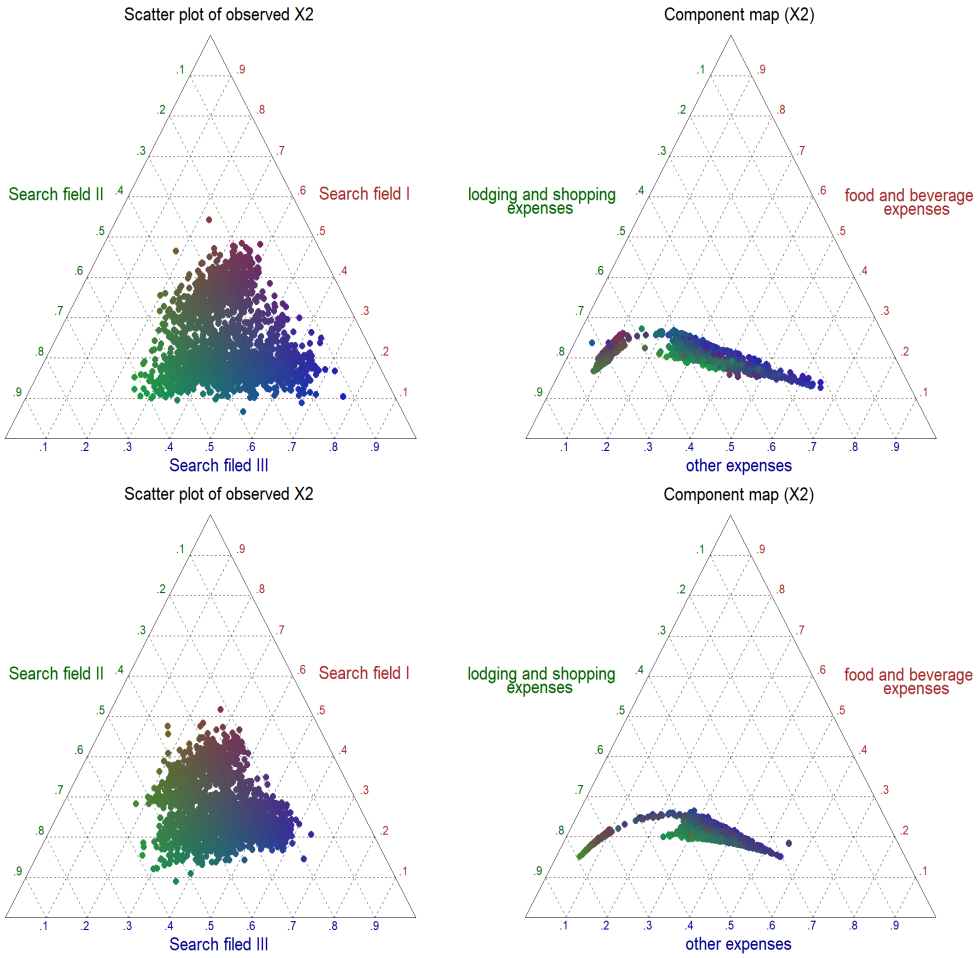

Figure 6: *Ternary plots of* $\boldsymbol{X}_2^a$ *(left) and* $\hat{\mathbf{f}}_2(\boldsymbol{X}_2^a)$ *(right) for the years 2019 (top) and 2021 (bottom) in the* 2^{nd} *cluster.*

1) 군집 1 회귀분석 결과

Figure 3은 2019년과 2021년 자료의 $\boldsymbol{X}_1^a$와 $\hat{\mathbf{f}}_1(\boldsymbol{X}_1^a)$를 나타낸 것이다. 왼쪽 삼각 도표에 있는 $\boldsymbol{X}_1^a$의 산점도를 보면, 두 연도 모두 Age I(20-30대)과 Age II(50대 이상)의 비율이 Age III(기타 연령대)보다 대체로 높은 것을 볼 수 있다. 오른쪽 삼각 도표에 있는 $\hat{\mathbf{f}}_1(\boldsymbol{X}_1^a)$를 보면, 점들이 COVID-19 이후보다 COVID-19 이전에 더 퍼져 있어 COVID-19 이전에 연령대별 방문자 수 비율이 관광지출액 비율에 미치는 영향이 더 컸고, 특별히 COVID-19 이전에는 그 영향이 식음료업과 기타업종에 있었는데, COVID-19 이후에는 이것이 숙박업 및 쇼핑업과 기타업종으로 옮겨간 것을 볼 수 있다.

Figure 4는 2019년과 2021년 자료의 $\boldsymbol{X}_2^a$와 $\hat{\mathbf{f}}_2(\boldsymbol{X}_2^a)$를 나타낸 것이다. 왼쪽 삼각 도표를 보면 검색 건수 비율이 주로 Search field I(음식)에 몰려있고, COVID-19 이후에 이러한 경향이 좀 더 심화된 것을 볼 수 있다. 오른쪽 삼각 도표를 보면, 검색 건수 비율은 숙박업 및 쇼핑업 지출액과 기타업종 지출액에 큰 영향을 주지만, 식음료업 지출액에는 큰 영향을 주지 않는 것으로 드러났다. 그리고 이러한 경향은 COVID-19 이후에 좀 더 두드러졌다.

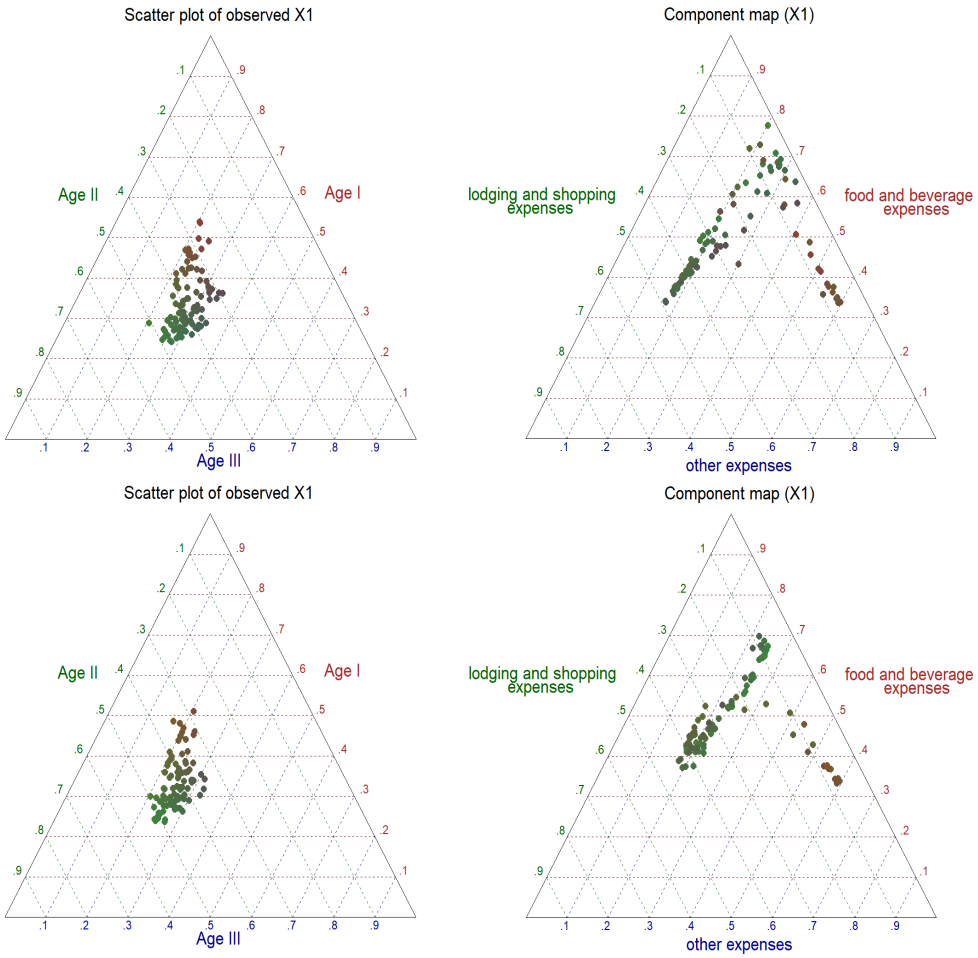

Figure 7: *Ternary plots of* $\boldsymbol{X}_1^a$ *(left) and* $\hat{\mathbf{f}}_1(\boldsymbol{X}_1^a)$ *(right) for the years 2019 (top) and 2021 (bottom) in the* 3^{rd} *cluster.*

2) 군집 2 회귀분석 결과

Figure 5는 2019년과 2021년 자료의 $\boldsymbol{X}_1^a$와 $\hat{\mathbf{f}}_1(\boldsymbol{X}_1^a)$를 나타낸 것이다. 왼쪽 삼각 도표를 보면 두 연도 모두 대체로 Age II(50대 이상), Age I(20-30대), Age III(기타 연령대) 순서로 방문자 수 비율을 차지하고 있다. 특별히 COVID-19 이전에는 Age I(20-30대)의 비율이 높을수록 식음료업에 지출하는 비율이 증가하고, Age II(50대 이상)의 비율이 높을수록 기타업종에 지출하는 비율이 증가하는 경향을 볼 수 있다. COVID-19 이후에는 Age I(20-30대)의 비율이 높을수록 식음료업 대신 숙박업 및 쇼핑업에 지출하는 비율이 증가했고, Age II(50대 이상)의 비율이 높을수록 기타업종에 이전보다 더 지출하는 경향을 보였다.

Figure 6은 2019년과 2021년 자료의 $\boldsymbol{X}_2^a$와 $\hat{\mathbf{f}}_2(\boldsymbol{X}_2^a)$를 나타낸 것이다. 왼쪽 삼각 도표를 보면 검색 건수 비율이 군집 1에 비해 넓게 퍼져 있어 더 다양하게 관광 유형을 검색하는 것을 알 수 있다. 오른쪽 삼각 도표를 보면, Search field II(숙박과 쇼핑)와 Search field III(기타 유형)의 검색 건수 비율이 높을수록 각각 숙박업 및 쇼핑업 지출 비율과 기타업종 지출 비율이 증가하는 경향이 있으나, 분야별 검색 비율이 식음료업 지출 비율에는 뚜렷한 영향을 주지 않는 것으로 드러났다. 그리고 이러한 경향성은 COVID-19 이후에도 큰

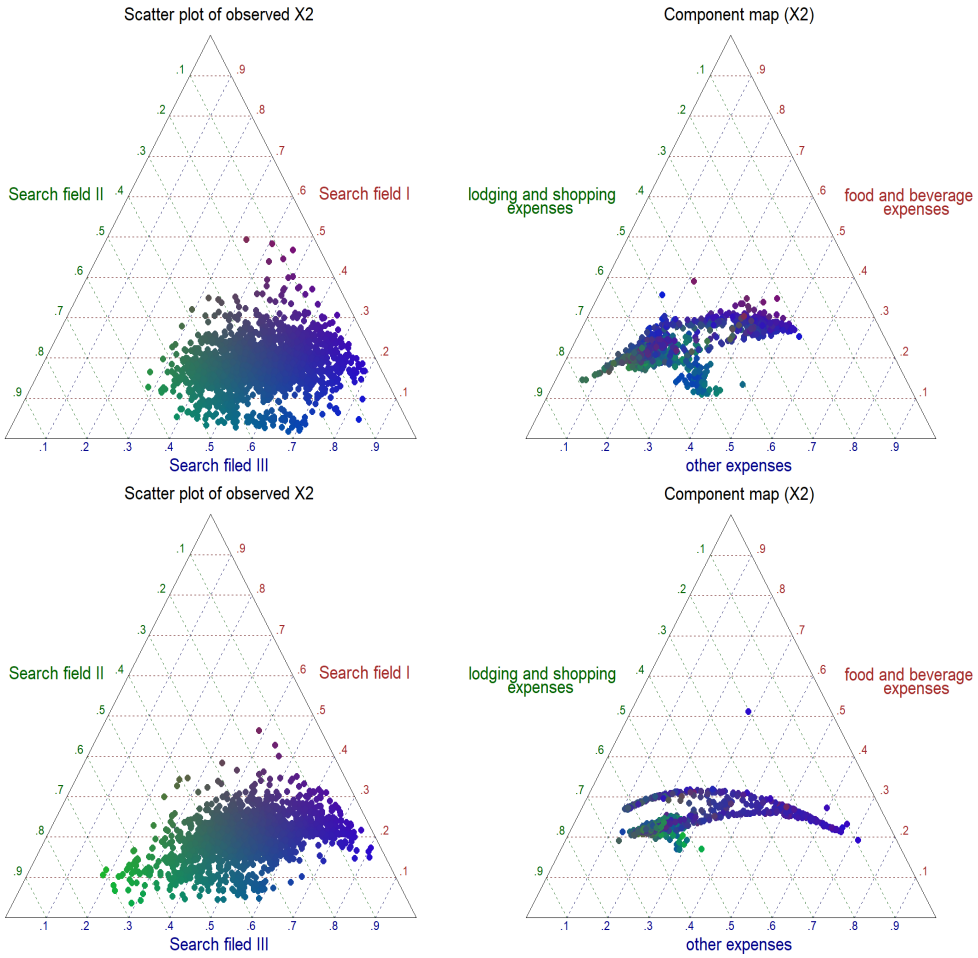

Figure 8: *Ternary plots of* $\boldsymbol{X}_2^a$ *(left) and* $\hat{\mathbf{f}}_2(\boldsymbol{X}_2^a)$ *(right) for the years 2019 (top) and 2021 (bottom) in the* 3^{rd} *cluster.*

변화없이 나타났다.

3) 군집 3 회귀분석 결과

Figure 7은 2019년과 2021년 자료의 $\boldsymbol{X}_1^a$와 $\hat{\mathbf{f}}_1(\boldsymbol{X}_1^a)$를 나타낸 것이다. 왼쪽 삼각 도표를 보면 두 연도 모두 Age I(20-30대)과 Age II(50대 이상)가 Age III(기타 연령대)에 비해 전반적으로 높은 비율을 차지하는 것을 볼 수 있다. 오른쪽 삼각 도표를 보면 연령대별 방문자 수 비율이 관광지출액 비율에 상당한 영향을 주지만, 그 경향성은 뚜렷하지 않다.

Figure 8은 2019년과 2021년 자료의 $\boldsymbol{X}_2^a$와 $\hat{\mathbf{f}}_2(\boldsymbol{X}_2^a)$를 나타낸 것이다. 왼쪽 삼각 도표를 보면 두 연도 모두 다른 군집에 비해 검색이 Search field II와 III에 집중된 경향을 보였다. 오른쪽 삼각 도표를 보면 Search field II(숙박과 쇼핑)와 Search field III(기타 유형)의 검색 비율이 높을수록 각각 숙박업 및 쇼핑업 지출 비율과 기타업종 지출 비율이 증가하는 경향이 있으나, 식음료업 지출 비율이 다소 감소하는 경우가 존재하였다. 이러한 경향은 군집 2와 비슷하지만, 군집 2보다는 덜 뚜렷하였다. 특별히 COVID-19 이후에 분야별 검색 비율에 따른 기타업종 지출 비율이 더 크게 차이났다.

5. 결론

본 연구는 강원도를 방문한 외지인의 방문추이에 따라 지역군을 나누고, 각 지역군마다 관광 특성을 파악할 수 있는 모형을 추정하고, 추정된 모형을 통해 COVID-19 이후 관광 특성의 변화에 대해 살펴보고자 하였다. 외지인의 방문추이에 따라 강원도는 행정도시 지역군(군집 1), 유명관광도시 지역군(군집 2)과 그외 소도시 지역군(군집 3)으로 구분할 수 있었다. 각 지역군마다 방문자의 연령대 비율과 방문지에 대한 관심도를 나타내는 검색지 유형 비율에 따라 업종별 관광지출액 비율이 다름을 확인하였다. 특히 COVID-19의 영향으로 각 지역군의 관광 특성이 다소 변화한 것을 알 수 있었다. 이를 바탕으로 강원도의 각 지자체는 도시 유형에 따라 관광객의 특징을 파악하고, 업종별 관광지출액 분석을 통해 더욱 효과적인 관광 정책을 수립할 수 있을 것이라 기대한다.

References

Badr HS, Du H, Marshall M, Dong E, Squire MM, and Gardner LM (2020). Association between mobility patterns and COVID-19 transmission in the USA: A mathematical modeling study, *The Lancet Infectious Diseases*, **20**, 1247–1254.

Egozcue JJ, Pawlowsky-Glahn V, Mateu-Figueras G, and Barceló-Vidal C (2003). Isometric logratio transformations for compositional data analysis, *Mathematical Geology*, **35**, 279–300.

Jeon JM and Park BU (2020). Additive regression with Hilbertian responses, *Annals of Statistics*, **48**, 2671–2697.

Jeon JM, Park BU, and Van Keilegom I (2021). Additive regression for non-Euclidean responses and predictors, *Annals of Statistics*, **49**, 2611–2641.

Mammen E, Linton OB, and Nielsen JP (1999). The existence and asymptotic properties of a backfitting projection algorithm under weak conditions, *Annals of Statistics*, **27**, 1443–1490.

Pawlowsky-Glahn V and Buccianti A (2011). *Compositional Data Analysis : Theory and Applications*, John Wiley & Sons, Hoboken, NJ.

UNCTAD (2021). Global economy could lose over 4 trillion due to COVID-19 impact on tourism. Retrieved Jun. 30, 2021, Available from: https://unctad.org/news/global-economy-could-lose-over-4-trillion-due-covid-19-impact-tourism

Received March 12, 2023; Revised April 19, 2023; Accepted May 4, 2023

강원도 관광산업에 대한 비모수적 구성비 자료 분석

박성은[a], 전정민[b], 이영경[1,a]

[a]강원대학교 통계학과; [b]서울대학교 통계학과

요 약

국내 대표 관광지인 강원도는 관광수요가 일부 지역에만 편중되어 있어 지역별로 다른 관광추이를 보인다. 따라서 각 지자체의 관광 활성화 방안 수립과 지역간 균형발전을 위해 각 지역의 관광 특성을 파악하고, 연도별 관광패턴을 비교하는 것이 중요하다. 본 논문에서는 최근 4년간의 강원도 관광 자료를 이용하여 지역을 군집화하고, 군집별 관광패턴을 Jeon 등 (2021)이 제안한 비유클리디안 가법모형으로 분석하였다. 이때, 연령대에 따른 방문자 수 비율과 방문지 유형에 따른 내비게이션 검색 수 비율을 공변량으로 하고, 업종별로 구분된 관광지출액 비율을 반응 변수로 하였다. 모형의 추정을 위해 평활역접합 방법과 성분별 띠폭 선택법을 이용하였다. 그리고 삼각 도표를 통해 추정된 모형을 시각화하고, 군집별로 적합 오차에 대한 예측 오차 비율을 비교하여 연도별 관광패턴 변화를 확인하였다.

주요용어: 구성비 자료, 성분별 띠폭 선택법, 비유클리디안 가법모형, 평활역적합, 심플렉스

이 논문은 한국연구재단 중견연구사업의 지원을 받아 수행된 연구임 (NRF-2021R1A2C1003920).

[1]교신저자: (24341) 강원도 춘천시 강원대학길 1, 강원대학교 통계학과. E-mail: youngklee@kangwon.ac.kr