Annals of Statistical Research
Vol. 5. No. 1, 2009, pp.1~17

A Comparison and Application of the Clustering Methods

Sung-Joon Hong[1] Ji-Ae Song[2] Su-Jung Jung[2] Young-Jin Jung[2]
Eun-Young Kim[2] Hyun-Jung Noh[2] Yong-Sung Joo[3]

Abstract

Many clustering analysis methods have been developed and used in various research area. In this paper, many well-known clustering analysis methods are applied and compared to simulated and actual data set.

Key Words: clustering, comparison, K-means, K-medoids, Fuzzy, hierarchical, SOM, SOTA, bagged, spectral

1) Corresponding Author : PhD. Student, Department of Statistics, Dongguk University, Pildong 3-26, Joonggu, Seoul, Korea. 100-715, E-mail : hsj8129@dongguk.edu
2) Graduate student, Department of Statistics, Dongguk University, Pildong 3-26, Joonggu, Seoul, Korea. 100-715, E-mail : sdaemon@dongguk.edu
2) Graduate student, Department of Statistics, Dongguk University, Pildong 3-26, Joonggu, Seoul, Korea. 100-715, E-mail : sj1004j@dongguk.edu
2) Graduate student, Department of Statistics, Dongguk University, Pildong 3-26, Joonggu, Seoul, Korea. 100-715, E-mail : jungjin0124@hanmail.net
2) Graduate student, Department of Statistics, Dongguk University, Pildong 3-26, Joonggu, Seoul, Korea. 100-715, E-mail : key0223@dongguk.edu
2) Graduate student, Department of Statistics, Dongguk University, Pildong 3-26, Joonggu, Seoul, Korea. 100-715, E-mail : hjhj2143@dongguk.edu
3) Assistant Professor, Department of Statistics, Dongguk University, Pildong 3-26, Joonggu, Seoul, Korea. 100-715, E-mail : yongsungjoo@dongguk.edu

Ⅰ. 서 론

군집이란 자연적으로 이루어지는 서로 상호간의 유사성을 가진 집단을 의미하며, 군집분석은 이미 정해져있는 군집을 알지 못할 때, 각 객체들 사의의 유사성을 측정하여 적절하게 군집을 형성하는 일련의 통계적 분석방법이다. 최근까지 많은 군집분석방법들이 개발되어 왔는데, 그 응용에 있어서 과연 어떤 방법들이 일반적으로 '좋은'군집화와 '나쁜'군집화를 이루어내는지 알아보는 것이 중요하다. 본 논문에서는 일반적으로 많이 사용되는 여러 군집분석 방법들에 대해 알아보고, 모의실험과 실제 데이터를 이용한 분석을 통하여 일반적으로 '좋은'군집화를 이루어내는 방법들과 '나쁜'군집화를 이루어내는 방법들에 대해 살펴본다.

Ⅱ. 군집 분석 방법

이 절에서 우리는 여러 가지의 주요한 군집분석방법들을 소개하고, 그 장·단점 및 특징들을 알아보고자 한다.

1. K-means

K-means(MacQueen(1967))는 군집의 수를 미리 결정하고, 중심점과 주어진 객체의 거리를 계산하여 가장 가까운 중심에 주어진 객체를 할당하는 방법이다.

먼저 k개의 초기 중심점들(군집들의 개수)을 선택한 다음, 각각의 객체는 가장 가까운 중심점에 할당되며, 각 중심점에 할당된 객체들의 집합이 군집이 된다. 이 때, 각 군집들의 중심점은 군집에 할당된 객체들을 기반으로 하여 갱신되며, 이러한 할당과 갱신 단계를 반복하여 어떤 객체도 군집이 바뀌지 않거나 또는 중심점들이 동일하게 유지될 때까지 계속 수행한다. 이와 같이 모든 객체들이 서로 중복되지 않는 군집들로 분할되는 것을 분할 군집(partitional clustering)이라 한다.

K-means는 중심점을 평균으로 계산하기 때문에, 평균을 구할 수 있는 데이터에서만 사용할 수 있다. 또한, 잡음과 이상치(outlier)에 대하여 민감한 결과를 보이며 초기에 선택한 K값이나, 비유사성을 계산하는 방법에 따라 다른 결과를 초래할 수 있다.

본 논문에서는 최적의 군집 개수를 결정하기 위한 방법들 중 하나로 Silhouette Width를 사용한다. Silhouette Width는 모든 객체들의 Silhouette value의 평균값이다. i번째 객체의 Silhouette value는 다음과 같이 계산 할 수 있다.

$$S(i) = \frac{b_i - a_i}{\max(b_i, a_i)}$$

여기서, a_i는 i번째 객체와 i번째 객체가 속한 군집 내에 있는 다른 모든 객체와의 거리들의 평균을 나타내며, b_i는 i번째 객체와 i번째 객체가 속한 군집과 가장 가까운 군집 내에 있는 다른 모든 객체와의 거리들의 평균을 나타낸다. Silhouette Width는 -1에서 1까지의 값을 가질 수 있으며, 가장 큰 값을 가질 때의 군집의 개수가 최적인 것으로 결정한다 (Julia Handl(2005)).

2. K-medoids

K-medoids는 K-means와 관련된 방법으로써, K-menas에서는 군집의 중심으로 객체들 간의 평균을 바탕으로 찾아낸 새로운 객체를 사용하는 것과 달리 군집의 중심으로 실제 객체인 메도이드(medoid)를 중심으로 사용한다. 여기서 메도이드란 “군집 내에서 가장 중심에 위치한 객체”(Kaufman, Rousseeuw(1990))를 말한다. K-medoids는 중심점을 메도이드를 사용하기 때문에, K-means의 중심점인 평균보다 이상치에 민감하지 않으므로 데이터들을 효과적으로 다룰 수 있다.

일반적으로 잘 알려진 K-medoids 방법들 중에는 PAM과 CLARA가 있다. Kaufman과 Rousseeuw(1978)에 의해 제안된 PAM은 먼저 k개의 초기 중심점들(군집들의 개수)을 선택한 다음, 각각의 객체는 가장 가까운 중심점에 할당되며, 각 중심점에 할당된 객체들의 집합이 군집이 된다. 이상적인 메도이드를 찾기 위해서는 반복을 통하여 메도이드들을 변화시켜 나가는데, 이렇게 메도이드들을 변화시킬 때마다 객체들이 가까운 메도이드들을 중심으로 객체를 형성하기 위해 움직이게 된다. 이때, 객체들이 변화된 메도이드로 인하여 재분류 되었을때 이동한 객체와 본래의 메도이드, 변화된 메도이드 간의 거리의 차를 비용이라고 한다. 만약 총 비용이 음수라면, 비용을 줄일 수 있는 메도이드로 바꾸어야 하고, 비용이 양수라면, 현재의 메도이드가 가장 적합한 메도이드임을 나타낸다. 하지만 데이터의 크기가 커질수록 계산량이 많아 컴퓨터의 수행속도가 느려진다(Han, Kamber(2000)).

이러한 PAM의 단점을 보안하기 위해 Kaufman과 Rousseeuw(1986)에 의해 제안된 CLARA가 있다. CLARA는 전체 데이터에서 표본을 랜덤하게 추출한 다음, 추출한 랜덤 표본에 PAM 방법을 적용시켜 메도이드를 찾는 과정을 반복하여 최적의 군집을 찾는 방법이다. 하지만 CLARA는 sample size에 의존하며, 만약 추출된 표본에서 좋은 메도이드가 없다면 최적의 군집을 찾지 못하는 단점이 있다.

K-medoids는 K-means와 마찬가지로 분할 군집(partitional clustering)이므로 최적의 군집의 개수를 결정하기 위해서 Silhouette Width를 흔히 사용한다.

3. Fuzzy Clustering

Fuzzy Clustering(Bezdek(1976))은 각각의 객체가 각 군집에 속하는 정도를 나타내는 가중치를 부여하여 조금 더 큰 가중치를 가지는 군집에 객체를 할당하는 방법이다. 본 논문에서는 Fuzzy Clustering의 방법들 중 c-means를 사용하였다. Fuzzy c-means는 Fuzzy 논리와 Fuzzy집합 이론의 개념을 사용하여 K-means와 유사하고 일반적인 방법으로 해석될 수 있지만, K-means는 어느 한 군집에 소속인가 아닌가를 보여주는 반면, Fuzzy c-means는 각 객체가 군집에 속하는 정도(degree of membership)를 제공한다는 차이가 있다(Pang-Ning Tan et al(2005)).

Fuzzy c-means는 객체들과 군집의 중심점의 거리를 최소화하기 위하여 객체들이 한 군집에서 다른 군집으로 이동할 수 있는 유동성을 가지고 있지만, 계산량이 많아지고 K-means 와 마찬가지로 초기에 선택한 값에 따라 다른 결과를 초래하기도 한다(Bezdek(1976)).

Fuzzy c-menas 역시 분할 군집(partitional clustering)이므로 최적의 군집 개수를 결정하기 위해 Silhouette Width를 흔히 사용한다.

4. 계층적 군집방법(Hierarchical Clustering)

계층적 군집방법은 사전에 군집 개수를 결정하지 않고 단계적으로 서로 다른 군집 결과를 제공하는 것으로써 병합형과 분할형으로 나누어진다. 병합형은 각 객체들을 별개의 군집으로 시작하여 각 단계마다 가장 가까운 군집들을 합하는 방법이며, 분할형은 모든 객체들을 하나의 군집으로 시작하여 각 객체들이 단일 원소 군집이 될 때까지 군집을 나누는 방법이다(Pang-Ning Tan et al(2005)). 일반적으로 병합형이 많이 쓰이며, 본 논문에서도 병합형에 대해서만 고려하였다.

병합형 계층적 군집방법에서는 군집간의 거리를 정의하는 방법에 따라서 <표 1>과 같이 구분된다.

<표 1> 군집 간 거리 정의에 따른 병합형 계층 군집방법

방법	군집 간 거리 정의
최단연결법	각기 다른 군집에 속한 가장 가까운 두 객체 사이의 거리
최장연결법	각기 다른 군집에 속한 가장 먼 두 객체 사이의 거리
평균연결법	각기 다른 군집에 속한 모든 객체들의 쌍 간에 평균 거리
중심연결법	각 군집에 속한 객체들의 평균점 사이의 거리
워드 방법	두 군집이 병합될 때 오차의 제곱의 증가분

계층적 군집방법은 객체가 어떤 군집에 할당이 되면 다른 군집으로는 다시 할당될 수 없으며, 대용량의 데이터에는 적절하지 않을 수도 있다.(김기영, 전명식(1990))

계층적 군집방법의 최적의 군집 개수를 결정하기 위해 R-square와 Pseudo F, Pseudo T^2 통계량을 확인해야 한다. R^2는 군집 집단의 변동에 의하여 설명되는 비율로써 군집간 제곱합을 총 제곱합으로 나누어서 계산할 수 있다(최용석, 정광모(2001)). 어느 단계에서 R^2가 급격히 감소하면 그 전 단계가 최적 군집 개수로 결정된다. Pseudo F 통계량은 군집 간 평균 제곱합을 군집 내 평균 제곱합으로 나눈 값으로 이 값 역시 어떤 단계에서 두 군집이 결합되면서 급격히 감소되면 그 이전 단계가 최적의 군집 개수가 된다. Pseudo T^2 통계량은 다음과 같이 계산되어진다.

$$T^{*2} = \frac{[SSW_t - SSW_r - SSW_s](n_r + n_s - 2)}{SSW_r + SSW_s}$$

여기서 SSW_t는 전체 자료의 군집 내 제곱합을 의미하며, SSW_r와 SSW_s는 각 군집의 군집 내 제곱합, n_r과 n_s는 각 군집의 크기이다. 병합이 되면서 Pseudo T^2 통계량이 다른 단계와 비교하여 유의하게 크다면 그 전 단계가 최적의 군집 개수로 결정된다(이성석(1997)).

5. SOM(Self-Organizing Maps)

SOM(Kohonen(2001))은 신경망 모형을 기반으로 한 군집분석 방법이다. 다차원의 데이터를 2차원 평면공간에 표현하는 차원 축소에 많이 이용된다. 입력벡터(input vector: 입력 데이터)와 가장 유사한 노드(node)로 채택된 하나의 노드를 중심으로 이웃반경(neighborhood radius) 내에 있는 노드들만의 학습을 할 수 있는 승자독점(winner take

all) 학습철학을 채택한다.

SOM의 구조는 입력데이터의 개수만큼의 노드들을 가진 입력층(input layer)과 군집의 수만큼의 뉴런들을 가진 경쟁층(competive layer) 2개의 층으로 구성된다. 입력층에서 경쟁층 방향으로 모든 노드간에는 완전한 연결이 이루어져 있다. 그리고 경쟁층 각 노드간에 고밀도, 완전연결(full connected) 격자(grid)로 구성되어 되어 있다(김대수(1992)).

SOM은 먼저 군집의 개수 k, 이웃반경 $h(\vec{y}, \sigma)$ 그리고 학습비율 $\epsilon(t)$의 값을 지정하고 경쟁층의 각 뉴런들의 초기 연결 가중치 벡터 $\vec{w}$를 0과 1사이의 값으로 생성한다. 그리고 경쟁층 위에 지정된 군집의 개수를 지정한 연결 가중치를 기준으로 노드를 놓고 모든 입력변수들의 값을 0과 1사이의 값으로 변형 시킨 다음 경쟁층에 있는 노드과 입력벡터와의 유클리디안 거리(Euclidean distance)에 의해 가장 유사한 연결 가중치를 가지는 노드(BMU : best matching unit)을 선택하여 그 노드을 기준으로 초기에 주어진 이웃반경에 안에 있는 모든 노드들의 연결 가중치를 갱신(update)한다.

갱신된 가중치 만큼 경쟁층에 있는 노드들의 위치가 이동하게 되는데 이 과정을 학습(training)이라고 하며, 학습과정을 반복해 나가면서 경쟁층의 있는 각 노드들의 가중치는 계속 갱신된다. 이 때, 이웃반경, 학습비율의 크기는 줄어들어 0에 가깝게 된다. 일정한 횟수의 반복된 학습 과정 후에 경쟁층의 있는 노드들은 입력데이터의 정보에 의해 격자형태로 고정되고, 그곳에 입력데이터를 투영하여 군집을 형성하게 된다(Kohonen(2001)).

SOM은 구조상의 다른 군집방법 보다 수행이 상당히 빠른 모형이며, 연속적인 학습이 가능하며 네트워크의 크기가 클수록 잘 작동한다. 반면에 군집의 개수를 정하기 어렵고 유일한 결과를 보장하지 않으며 또한 계층적인 구조(tree)로 표현할 수 없다(김대수(1992)).

SOM 역시 분할 군집(partitional clustering)이므로 최적의 군집개수 결정방법으로 흔히 Silhouette Width를 사용한다.

6. SOTA(Self-Organizing Tree Algorithm)

SOTA는 Hierarchical Clustering과 SOM의 장점을 결합한 방법이다. 먼저 초기 노드값을 랜덤(random)하게 선택하여 두 개의 셀(cell)로 나눈 다음, 새로운 자료들을 모든 셀과 비교하여 가장 유사한 셀에 할당을 한다. 각 셀에 할당된 셀의 분산을 계산하여 분산이 큰 셀을 다시 가지를 뻗어나가는 형태를 가진다. SOTA에서 가지는 자신이 원하는 군집 수가 될 때까지 가지를 뻗을 수 있으며(Javier Herrero(2001)), 해당 셀에 자료가 1개이거나 자료가 유사하다면 가지를 뻗는 것을 멈추게 된다.

SOTA는 Hierarchical Clustering과 SOM 보다 이상치에 더 로버스트한 편이며, 수행시간도 빠르다(Longed et al(2006)).

SOTA 역시 분할 군집(partitional clustering)이므로 최적의 군집 개수를 결정하기 위해 Silhouette Width를 흔히 사용한다.

7. Bagged Clustering

Bagged Clustering은 분할 군집방법 (partitional Clustering)과 Hierarchical Clustering (계층적 군집방법)이 결합된 방법이다. 새로운 부스트랩 데이터를 반복적으로 형성하며 군집화 하는 방법으로 데이터 마이닝 기법 중 앙상블 기법인 Bagging에 기반을 두고 있다 (Friedich Leisch(1999)). Bagged Clustering은 원래의 데이터로부터 B개의 부스트랩 샘플을 생성한 다음 각각의 부스트랩 샘플에 대해 K-means을 이용하여 군집분석을 실시한다. K-means를 이용한 결과 하나의 부스트랩 샘플에 대해 k개의 군집 중심점이 생성되고 총 k×B개의 군집 중심점을 형성하게 된다. 그런 다음 모든 군집 중심점을 결합하여 새로운 데이터 셋을 형성하고 이 데이터 셋에 대해 Hierarchical clustering을 이용하여 군집화한 후, 군집화된 군집 중심점에 대해 원래의 데이터를 할당하여 군집화하는 방법이다.

Bagged clustering은 하나의 데이터 셋으로부터 형성된 부스트랩 데이터에 대해 반복적으로 K-means를 수행함으로써 초기치에 민감한 K-means의 단점을 보완한 방법으로 생각할 수 있다(Friedich Leisch(1999)).

Baggged clustering에서 최적의 군집 개수는 K-means에 의해 생성된 k×B개의 군집 중심점들에 대한 덴드로그램에 의해 결정된다.

8. Spectral Clustering

Spectral Clustering은 컨벡스(convex)형태가 아닌 군집들을 찾아낼 수 있는 기법이다. 컨벡스 셋은 어떤 군집 안에 있는 임의의 객체 2개를 연결할 때, 이 연결선이 군집 안에 위치하고 있는 경우를 말한다. 즉, Spectral Clustering은 이런 컨벡스 셋 형태를 이루지 않는 데이터에서 군집을 분류하기에 적합한 방법이다.

Spectral Clustering은 분포에 영향을 받지 않으며, 고유값 일부를 사용함으로서 일부(local)정보를 바탕으로 전체적인(global) 군집을 할당한다(Francis et al(2003)). 군집의 개수에 따라 유사도 행렬(affinity matrix)을 만드는 방법은 달라지게 되는데, 먼저 군집을 두

개로 나눌 때에는 라플라시안(Laplacian)행렬을 만든다. L(Laplacian)=D-A의 형태로 $A = w_{ij},\ i,j = 1,\cdots,n$(affinity)행렬은 $n \times n$개의 객체들간의 근접도 가중치를 행렬형태로 나타난 것이다. 이 가중치는 가까이 있는 객체들에는 높게, 멀리 있는 객체들은 낮게 나타낸다. $D(i,i) = \sum_{j=1}^{n} w_{ij},\ i = 1,\cdots,n$(degree($n \times n$))행렬은 객체들의 가중치를 더한 대각행렬이고, L(Laplacian)=D-A 행렬에서 나온 두 번째 큰 고유값의 고유벡터를 이용하여 두 개의 군집으로 나누게 된다. 군집의 개수가 3개 이상일 경우는 L행렬 대신 A'행렬을 사용하는 $A' = D^{-1/2} A D^{-1/2}$이다(Pejus Das, Mathew Beal(2004)). A'행렬에서 나온 k번째 고유값에 해당하는 고유벡터를 이용하여 k개의 군집으로 나누게 된다(Urike von Luxburg(2006)).

Spectral Clustering은 수학적으로 정립이 잘 되어 있으며 구현이 간단하고 컨벡스 형태가 아닌 군집의 분류에 적당하다. 또한 대용량 자료에서도 효율적이다(Urike von Luxburg(2006)). 그러나, K-means나 Fuzzy clustering과는 달리 일부(local)정보만을 사용하는 한계가 있으며, 크기가 다른 군집들을 찾아내기 어렵다는 단점이 있다.

Spectral Clustering에서는 연속적인 두 개의 고유값의 차이를 계산하여 가장 차이가 큰 값($\max \Delta_k |\lambda_k - \lambda_{k-1}|$)을 최적의 군집개수로 결정한다.

Ⅲ. 군집분석 평가지표

외부 기준에 대해 군집 결과를 비교하기 위해서는 평가지표가 필요하다. 이 평가지표를 비교하여 최적의 군집 방법을 선택할 수 있다. 다음의 6가지 평가지표는 유사성 계수로써 0과 1사이의 값을 가지며 1에 가까울수록 군집이 잘 되었음을 의미한다.

동일한 데이터셋에 대해 U는 집단 표시, V는 군집 분석 결과를 나타낸다고 하자. a는 U에서 같은 집단 표시를 가지고, V에서도 같은 군집에 속하는 객체들의 쌍의 수, b는 U에서는 같은 집단 표시를 가지나, V에서는 다른 군집에 속하는 객체들의 쌍의 수, c는 V에서는 같은 군집에 속하나, U에서는 다른 집단 표시를 갖는 객체들의 쌍의 수 그리고 d는 V에서도 다른 군집에 속하고, U에서도 다른 집단 표시를 갖는 객체들의 쌍의 수라고 하면 <표 2>와 같은 결과를 얻을 수 있다. 여기서 1은 각 객체들이 동일한 집단이나 군집에 속하는 경우, 0은 그렇지 않은 경우를 나타낸다.

<표 2> U(집단표시)와 V(군집분석 결과)에 대한 교차표

U \ V	1	0	sum
1	a	b	a+ b
0	c	d	c+ d
sum	a+ c	b+ d	a+ b+ c+ d

<표 2>를 이용한 흔히 사용되는 5개의 평가지표들을 <표 3>에 정리하였다.

<표 3> 군집분석 평가지표

Validation Index	Formula
Rand Index (Simple Matching)	$\frac{a+d}{a+b+c+d}$
Double Matching	$\frac{2(a+d)}{2(a+d)+b+c}$
Roser-Tanimoto	$\frac{a+d}{a+d+2(b+c)}$
Rusell-Rao	$\frac{a}{a+b+c+d}$
Jaccard	$\frac{a}{a+b+c}$

Rand index는 모든 객체의 개수에 대한 U에서 같은 집단 표시를 가지고, V에서도 같은 군집에 속하는 쌍으로 이루어진 객체의 개수와 V에서도 다른 군집에 속하고, U에서도 다른 집단 표시를 갖는 쌍으로 이뤄진 객체의 개수의 합의 비율로써 의학 분야에서 많이 적용된다(Bhaba R. Sarker(1996)).

Jaccard는 Rand Index (Simple Matching)에서 V에서도 다른 군집에 속하고, U에서도 다른 집단 표시를 갖는 쌍으로 이뤄진 객체의 개수를 포함하는 cell d를 제외한 수치이며, 비대칭의 이항변수에 유용하다.

Roser-Tanimoto는 Jaccard의 확장된 개념으로

$$T = \frac{N_c}{N_a + N_b - N_c}$$

여기서, N_a : A에 있는 항목 수, N_b :B에 있는 항목수, N_c: 교집합에 있는 항목 수 이다.

A집단과 B집단이 있다고 하면, 전체 원소들과 두 집단에 동시에 들어가 있는 원소의 비율을 의미한다. Rogers-Tanimoto는 그룹과 군집 결과가 각기 다른 가중치를 가지는 경우나 positive matching 또는 negative matching의 수치가 큰 경우 적용할 수 있는 유사성 척도로써 Jaccard index의 확장이라 할 수 있다(Javad Sadri et al(2006)).

Rusell-Rao index는 모든 객체의 개수에 대한 U에서 같은 집단 표시를 가지고, V에서도 같은 군집에 속하는 쌍으로 이루어진 객체의 개수의 비율로써 Jaccard index와 대조적이다(Holmes Finch(2005)).

위의 방법들 중 일반적으로 많이 쓰이는 방법은 Rand Index이다. 그러나 이 지표의 문제점은 Rand Index의 기대값이 0이라는 것이다. 이를 보완하기 위해 adjusted Rand Index가 개발이 되었다.

n_{ij}는 집단 표시 u_i와 군집 v_j에 모두 속하는 객체들의 수, $n_{i.}$과 $n_{.j}$는 각각 집단 표시 u_i에 속하는 객체들의 수, 군집 v_j에 속하는 객체들의 수라 하면 <표 4>를 이용하여 adjusted Rand Index는 다음과 같이 계산할 수 있다.

<표 4> 집단표시 u_i와 군집 v_j를 비교하기 위한 교차표

Class \ Cluster	v_1	v_2	$\cdots$	v_C	Sums
u_1	n_{11}	n_{12}	$\cdots$	n_{1C}	$n_{1.}$
u_2	n_{21}	n_{22}	$\cdots$	n_{2C}	$n_{2.}$
					$\vdots$
$\vdots$	$\vdots$	$\vdots$	n_{ij}	$\vdots$	$n_{i.}$
					$\vdots$
u_R	n_{R1}	n_{R2}	$\cdots$	n_{RC}	$n_{R.}$
Sums	$n_{.1}$	$n_{.2}$	$\cdots$ $n_{.j}$ $\cdots$	$n_{.C}$	$n_{..}=n$

$$adjusted\ Rand\ Index = \frac{\Sigma_{i,j}\binom{n_{ij}}{2} - \left[\Sigma_i\binom{n_{i.}}{2}\Sigma_j\binom{n_{.j}}{2}\right]/\binom{n}{2}}{\frac{1}{2}\left[\Sigma_i\binom{n_{i.}}{2}+\Sigma_j\binom{n_{.j}}{2}\right] - \left[\Sigma_i\binom{n_{i.}}{2}\Sigma_j\binom{n_{.j}}{2}\right]/\binom{n}{2}}$$

본 논문에서는 군집분석 결과의 평가를 위하여 위에서 살펴본 6가지의 평가지표들을 모두 사용하였다.

Ⅳ. 모의실험을 통한 군집분석 방법들의 비교

모의실험에서 우리는 3개의 군집으로 이루어진 3변량 자료를 생성하였다. 이 때, 각각의 군집들은 다음과 같은 분포를 가진다.

$$\text{군집}1 : \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \sim MN\left(\begin{bmatrix} 0 \\ 3 \\ 6 \end{bmatrix}, \begin{bmatrix} 4\,0\,0 \\ 0\,4\,0 \\ 0\,0\,3 \end{bmatrix} \right), \quad \text{군집}2 : \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \sim MN\left(\begin{bmatrix} 6 \\ 0 \\ 3 \end{bmatrix}, \begin{bmatrix} 4\,0\,0 \\ 0\,4\,0 \\ 0\,0\,3 \end{bmatrix} \right),$$

$$\text{군집}3 : \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \sim MN\left(\begin{bmatrix} 3 \\ 6 \\ 0 \end{bmatrix}, \begin{bmatrix} 4\,0\,0 \\ 0\,4\,0 \\ 0\,0\,3 \end{bmatrix} \right)$$

각 변수들로부터 100개씩 난수를 생성하여 얻어진 총 300개의 데이터를 앞에서 살펴보았던 군집분석 방법들을 이용하여 군집분석을 실시하였다. 이 때, K-medoids는 데이터의 개수가 많이 않으므로 PAM방법을 사용하였고, 계층적 군집방법에서는 중심연결법과 비슷하지만 군집중심간의 거리에 가중값을 부여하는 워드방법을 사용하였다. 군집의 개수를 결정하기 위하여 각 군집분석 방법별로 2절에서 언급한 방법들을 사용하였고, 1,000번의 모의실험을 통하여 얻어진 평가지표들의 평균값은 다음과 같다. 여기서, k는 평균 군집개수, RI는 Rand Index, SS은 Sokal-Sneath, RT는 Roser-Tanimoto, RR은 Russel-Rao, J는 Jaccard 그리고 adj RI는 adjusted Rand Index를 나타낸다.

<표 5> 1,000번의 모의실험을 통한 평가지표들의 평균값

Method	k	RI	SS	RT	RR	J	adj RI
K-means	3.006	0.982	0.991	0.966	0.323	0.952	0.960
K-medoids	3	0.985	0.993	0.971	0.324	0.956	0.966
Fuzzy Clustering	3	0.986	0.993	0.973	0.324	0.960	0.969
Hierarchical Clustering	2.738	0.919	0.955	0.864	0.320	0.837	0.839
SOM	3	0.848	0.917	0.740	0.273	0.649	0.668
SOTA	3.097	0.909	0.952	0.835	0.287	0.763	0.796
Bagged Clustering	3.004	0.972	0.986	0.946	0.317	0.920	0.937
Spectral Clustering	2	0.762	0.864	0.618	0.326	0.581	0.544

<표 5>에서 각 군집방법들에 따라 군집의 개수는 다르지만, 평가지표들의 값을 비교해 보았을 때, Fuzzy Clustering이 전체적으로 좋은 결과를 나타냈고 K-means, K-medoids, Bagged Clustering의 결과도 좋은 편이다. 반면에, 계층적 군집방법(Hierarchical Clustering), SOM, SOTA, Spectral Clustering은 낮은 결과를 나타냈으며, 특히 Spectral Clustering이 가장 좋지 않은 결과를 나타냈다.

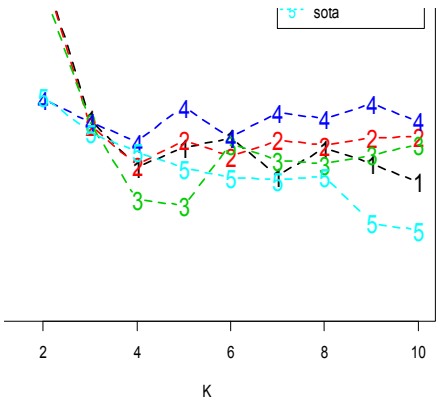
Ⅴ. 실제 사례 데이터를 이용한 군집분석방법들의 비교

수질이 농업활동에 미치는 영향을 이해하기 위하여 천층수, 암반수와 지표수에서 주요한 화학 성분들에 대한 조사가 한국에서 두 번째로 큰 낙동강 유역의 충적지역에서 수행되었다. 대체로 1년 내내 왕성한 농업활동이 이루어지고 있는 이 지역에는 많은 양의 인공 질소 화학비료와 복합 화학비료들이 농지에 사용되었고, 산성화된 토양을 중화시키기 위해 주로 석회(CaO)가 사용된다.

전체 연구 지역에서 화학물질들의 영향을 알아보기 위해 넓은 범위에 걸쳐 1999년 10월부터 2000년 9월까지 1년간 80곳의 물을 측정하였다.

화학비료와 관계된 주요 성분들($\log(HCO_3^-)$, $\log(Ca^{2+})$, $\log(NO_3^-)$, $\log(SO_4^{2-})$, $\log(Cl^-)$)의 연평균 농도를 측정하였으며, 특히 NO_3^-와 Ca^{2+}는 비료와 석회의 주요한 성분들이다. 위 논문에서는 낙동강 유역의 천연수는 중성 pH를 가지므로 알칼리성은 물의 샘플들의 HCO_3^-의 농도와 일치한다.

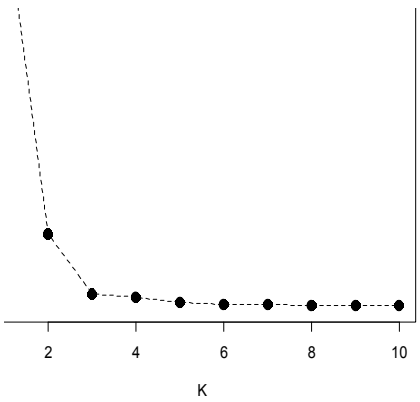
앞에서 살펴본 여러 군집분석 방법들을 이용하여 물의 샘플들의 군집을 분류해 보았다. 군집 분석에서는 군집의 개수를 결정하는 것이 중요한 문제이므로, 각 군집 방법 별로 군집 개수를 찾아보았다.



[그림 1] 분할 군집방법들의 Silhouette Width

[그림 1]은 K-means, K-medoids, Fuzzy clustering, SOTA와 SOM으로 군집분석을 하였을 때 군집의 개수에 따른 Silhouette Width를 나타낸 것이다. Silhouette Width가 가장 클 때를 군집개수로 하므로 [그림 1]에 나타난 5가지 군집 방법 모두 군집의 개수는 2로 나타났다.

[그림 2]은 Spectral Clustering에서 고유값의 크기에 따른 Scree Plot을 나타낸 것이다. 1과 2사이에서 고유값의 차이가 가장 크므로 군집개수는 2로 나타났다.



[그림 2] Spectral Clustering의 Scree Plot

계층적 군집방법과 Bagged Clustering은 덴드로그램에서 height값을 비교한 결과 군집개수가 각각 계층적 군집방법은 3, Bagged Clustering은 2로 나타났다.

위에서 얻어진 군집개수를 이용하여 각 방법별로 군집분석을 실시한 결과는 다음과 같다.

<표 6> 낙동강 자료를 이용한 평가지표들의 값

Method	k	RI	SS	RT	RR	J	adj RI
K-means	2	0.510	0.676	0.342	0.333	0.405	0.096
K-medoids	2	0.510	0.676	0.342	0.333	0.405	0.096
Fuzzy Clustering	2	0.541	0.702	0.370	0.331	0.419	0.142
Hierarchical Clustering	3	0.670	0.802	0.503	0.235	0.416	0.313
SOM	2	0.557	0.715	0.386	0.236	0.348	0.114
SOTA	2	0.689	0.816	0.525	0.321	0.508	0.386
Bagged Clustering	2	0.510	0.676	0.342	0.333	0.405	0.096
Spectral Clustering	2	0.510	0.676	0.342	0.333	0.405	0.096

<표 6>에 나타난 결과를 보면 K-means, K-medoids, Bagged Clustering, Spectral Clustering의 평가지표들의 값이 동일한 것을 알 수 있다. 각 군집방법들에 따라 군집의 개수는 다르지만, 대체적으로 2개의 군집이 형성되었으며 SOTA와 계층적 군집방법(Hierarchical Clustering)이 대체로 좋은 결과를 나타냈다. 반면에, 위의 두 방법을 제외한 다른 방법들은 낮은 결과를 나타냈다.

Ⅵ. 결론

본 연구는 모든 군집분석 방법들을 다룬 것은 아니지만, 현재 많이 사용되는 방법들을 모의실험과 실제 자료에 활용하여 비교하였다는 점에서 그 의미가 있다. 이번 연구에서는 1,000번의 모의실험을 통한 결과로 K-means, K-medoids, Fuzzy Clustering, Bagged Clustering이 좋은 결과를 보였고, 낙동강 수질에 대한 실제 자료에서는 SOTA와 계층적 군집방법(Hierarchical Clustering)이 좋은 결과를 보였다. 위의 결과로 보아 자료의 분포에 따라 군집분석의 결과가 달라짐을 알 수 있다. 따라서 자료의 분포에 맞는 적절한 군집분석 방법을 적용하는 것이 중요하며, 자료의 분포에 따라 어떠한 군집분석 방법이 좋은 결과를 나타내는지에 대한 추가적인 연구가 필요하다.

본 연구에서는 전통적인 군집분석 방법을 많이 다루었고, 모의실험을 위해 생성한 자료는 잡음을 포함하지 않았다는데 그 한계가 있다. 생성한 자료는 다변량 정규분포를 따르므로 Model based Clustering과 같은 군집분석방법도 고려할 수 있으며, 잡음(noise)을 포함한 자료를 이용하여 군집분석 방법을 비교하는 것 또한 의미가 있을 것이다.

참고문헌

[1] 김기영, 전명식 (1990). SAS 군집분석. 자유아카데미.

[2] 김대수 (1992). 신경망 이론과 응용, 하이테크 정보. Page 169-188.

[3] 이성석 (1997). 군집분석에서 군집갯수에 관한 연구. 응용과학연구, Vol.6 No.1.

[4] 임대혁 (2004). 새로운 Fuzzy 집락분석방법과 Simulation 기법에 관한 연구. 大韓經營情報學會. 제14권.

[5] 최용석, 정광모 (2001). SAS를 활용한 응용 다변량 자료분석. 교우사.

[6] Bezdek (1976). A physical interpretation of Fuzzy ISODATA.

[7] Bhaba R. Sarker, The resemblance coefficients in group technology: A survey and comparative study of relational metrics, Computers ind. Engng Vol. 30, No. 1(1996), 103-116

[8] Francis R. Bach and Michael I. Jordan (2003). Learning Spectral Clustering.

[9] Friedich Leisch (1999). Bagged Clustering. Technical report, SFB Adaptive Information Systems and Modelling in Economics and Management Science, Vienna University of Economics and Business Administration, http://www.ci.tuwien.ac.at/ ~leisch/papers/fl-techrep.html.

[10] Forgy, E. W. (1965) Cluster analysis of multivariate data: efficiency vs interpretability of classifications. Biometrics 21, 768-769.

[11] G. J. McLachlan, R. W. Bean and D. Peel (2001). A mixture model- based approach to the clustering of microarray expression data.

[12] Han. J. and Kamber. M. (2000). Data Mining : Concepts and Techniques, The Morgan Kaufman Series in Data Management Systems. Morgan Kaufman Publishers.

[13] Hartigan, J. A. and Wong, M. A. (1979). A K-means clustering algorithm. Applied Statistics 28, 100-108.

[14] Holmes Finch, Comparison of Distance Measures in Cluster Analysis with Dichotomous Data, Journal of Data Science 3(2005), 85-100

[15] Javad Sadri, Ching Y. Suen, Tine D. Bui, A New Clustering Method for Improving Plasticity and Stability in Handwritten Character Recognition Systems, Vol. 2, 2006, 1130-1133

[16] Javier Herrero, Alfonso Valencia and Joaquin Dopazo (2001). A hierarchical unsupervised growing neural network for clustering gene expression patterns.

[17] Julia Handl, Joshua Knowles and Douglas B. kell (2005). Supplementary material to computational cluster validation in part-genomic data analysis. Bioinformatics. Vol 00. No. 00. 1-3.

[18] Ka Yee Yeung, Wlter L. Ruzzo (2001). Details of the Adjusted Rand index and Clustering algorithms Supplement to the paper "An empirical study on Principal Component Analysis for clustering gene expression data" (to appear in Bioinformatics).

[19] Kaufman L. and Rousseeuw. P. J. (1990). Find Groups in Data : an Introduction to Cluster Analysis. John Wiley & Sons.

[20] Lloyd, S. P. (1957, 1982) Least squares quantization in PCM. Technical Note, Bell Laboratories. Published in 1982 in IEEE Transactions on Information Theory 28, 128-137.

[21] Longed Yin, Chun-Hsi Huang and Jun Ni (2006). Clustering of gene expression data :performance and similarity analysis.

[22] MacQueen. J. (1967). Some methods for classification and analysis of multivariate observation. Proc. 5th Berkeley Symp. Math. Statist. Porb., 1:pp 128 - 297.

[23] Pang-Ning Tan, Michael Steinbach, Vipin Kumar (2005). Introduction to data mining. Addison Wesley.

[24] Pejus Das and Mathew Beal (2004). Techniques for Spectral Clustering.

[25] Ping Ma, Cristian I. Castillo-Davis, Wenxuan Zhong and Jun S. Liu (2006). A data-driven clustering method for time cousrse gene expression data. Nucleic Acids Research. Vol 34. No.4. 1261-1269.

[26] T. Kohonen. (2001). Self-Organizing Maps. volume 30 of Springer Series in Information Sciences. Springer-Verlag.

[27] Ulrike von Luxburg (2006). A Tutorial on Spectral Clustering. Technical Report No. TR-149, 2006.

[투고일: 2009. 2. 2] [심사일: 2009. 2.10] [게재확정일: 2009. 2.15]

Annals of Statistical Research
Vol. 5. No. 1, 2009, pp.19~32

Comparison and Application of the Outlier Detection Methods

Sun Jung[1], Young-Je Woo[2], Hye-Rim Lee[3], Yong-Sung Joo[4]

Abstract

The outlier detection is important process before it analyze. Because outliers have masking and swamping effects when multivariate data is analyzed. In this paper, we compare Mahalanobis distance, Hadi method and Local Outlier Factor method through simulation study also these three methods are applied to Seawater intrusion data set. In terms of correct detection rate of outliers, when the number of observations is large, Hadi method is better than the other method. When the number of outliers is relatively large, Hadi and LOF methods worked better than Mahalanobis distance.

Key Words: outlier, Mahalanobis distance, Hadi, LOF

1) Corresponding Author : PhD. Student, Department of Statistics, Dongguk University, Pildong 3-26, Joonggu, Seoul, Korea. 100-715, E-mail : iamlucky@dongguk.edu
2) Graduate student, Department of Statistics, Dongguk University, Pildong 3-26, Joonggu, Seoul, Korea. 100-715, E-mail : youngjw@dongguk.edu
3) Graduate student, Department of Statistics, Dongguk University, Pildong 3-26, Joonggu, Seoul, Korea. 100-715, E-mail : leehr26@dongguk.edu
4) Assistant Professor, Department of Statistics, Dongguk University, Pildong 3-26, Joonggu, Seoul, Korea. 100-715, E-mail : yongsungjoo@dongguk.edu

Ⅰ. 서 론

이상점(理想點, outlier)은 조사의 대상이 되는 모집단에 속하지 않는다고 의심이 될 정도로 정상범위 밖으로 아주 동떨어진 관찰점(한국통계학회, 통계용어사전 p.203)으로 정의되어진다. 주어진 자료에 대해서 올바른 결론을 도출하기 위해서는 이러한 이상점들의 존재 유무를 확인해야 한다. 왜냐하면 이런 이상점들은 편향된 추정량(biased estimator)을 제공하여 올바르지 않은 분석 결과를 가져다주기 때문이다. 이상점의 탐색방법들은 이상점을 제대로 찾아내는 "정탐률"과 이상점이 아닌 자료를 이상점으로 찾아내는 "오탐률" 두 가지 관점에서 평가할 수 있다. 정탐률은 높을수록, 오탐률은 낮을수록 좋다. 본 논문에서는 모의실험과 실제 자료를 이용한 분석을 통하여 각 이상점 탐색 방법들의 정탐율과 오탐률를 비교하고자 한다.

이상점들은 자료를 오염시키는 오류(error) 또는 잡은(noise)로 생각될 수 있으나 모두가 쓸모없는 것은 아니다. 때때로 이상점들은 우리가 인지하지 못하고 있던 중요한 정보를 담고 있기 때문에 우리가 주어진 자료에 대해서 올바른 분석 결과를 도출하길 원한다면 이상점 유무에 대한 탐색과, 발견된 이상점이 어떤 의미를 가졌는지 확인하는 작업이 매우 중요하다고 할 수 있겠다.

자료에서 이상점의 발생 원인을 다음과 같은 세 가지의 이유로 생각해 볼 수 있다.

첫 번째는 '오차(error)'이다. 오차라 함은 자료를 기록할 때, 부정확하게 기록됨으로 발생하는 것을 말한다. 두 번째로는 '오염(contamination)'이다. 이는 자료가 다른 모집단으로부터 발생된 것을 말한다. 세 번째는 Inherent variability이다. 이는 자료 자체가 갖는 변동의 폭이 큰 경우를 말한다.

관찰된 자료의 개수가 적은 경우에는 산점도 등을 통해서 쉽게 육안으로 이상점들을 확인할 수 있지만 자료의 수가 많거나(large data set), 변수의 수가 많은 경우(high dimensional multivariate data)에는 불가능하다.

이상점 탐색 방법들은 단변량 이상점 탐색법과 다변량 이상점 탐색방법으로 나눠서 생각할 수 있다. 초기에는 단변량 이상점 탐색법만으로 가능했으나, 최근의 대부분의 자료의 형태가 다변량이므로 다변량 이상점 탐색방법에 대한 연구가 많이 이루어지고 있다.

다른 관점으로 이상점 탐색 방법을 나눠보면 모수적(parametric) 방법과 비모수적방법(nonparametric)으로 생각해 볼 수 있다. 모수적(parametric) 방법은 단변량 자료일 때 주로 사용하는 방법으로 주어진 자료들이 어떤 분포를 따른다고 가정하고, 통계적 방법을 사용하여 모수를 추정한다. 이 방법은 가정한 분포로부터 벗어난 관측치들을 이상값(outlier)

로 판단한다. 비모수적 방법들은 데이터 마이닝 기법들을 사용하여 다변량 자료와 대용량 자료들에서 이상점을 탐색하는 방법들이다. 본 논문에서는 다변량 이상점 탐색방법에 대하여 모의실험과 사례 연구를 하고자 한다.

다변량 이상점 탐색방법을 사용할 때 이상점들의 가림효과(masking effect)와 대세효과(swamping effect)를 주의해야 한다. 가림효과란 이상점이 다른 이상점을 감추어 일반적인 자료로 보이게 하는 것으로 흔히 분류 방법(clustering method)을 이용하여 이상점의 유무를 판단할 때 주로 나타난다. 대세효과는 가림효과와는 반대의 개념으로 어떤 이상값이 다른 보통의 자료에 영향을 주어 보통의 자료를 이상값으로 보이게 하는 효과이다. (Acuna and Rodriguez, 2004)

Ⅱ. 이상점 탐색 방법

이 절에서 우리는 이상점 탐색의 여러 가지 방법들을 소개하고, 그것들의 장·단점 및 특징들을 알아보고자 한다.

1. Mahalanobis distance

마하라노비스 거리(Mahalanobis distance)는 1936년에 Maharanobis에 의해 개발된 거리 측도이다. 이 방법은 다른 패턴을 갖는 변수들 간에 상관 관계(correlation)를 이용한 방법으로 집단의 유사성(Similarity)을 결정하는데 유용한 방법이다.

어떤 평균 벡터 $\mu=(\mu_1,\mu_2,\mu_3,\cdots,\mu_p)^T$와 다변량 벡터 $x=(x_1,x_2,x_3,\cdots,x_p)^T$에 대한 공분산 행렬 $\sum$으로 부터의 마하라노비스 거리는 다음과 같이 정의된다.

$$D_M(x)=\sqrt{(x-\overline{\mu})^T\sum^{-1}(x-\overline{\mu})}$$

또한 마하라노비스 거리(Mahalanobis distance)는 같은 분포를 갖는 두 랜덤 벡터 $\vec{x}$, $\vec{y}$ 간에 비유사성의 측도로서 정의 될 수 있다.

$$d(\vec{x},\vec{y})=\sqrt{(\vec{x}-\vec{y})^T\sum^{-1}(\vec{x}-\vec{y})}$$

여기서 $\sum$은 공분산 행렬이다.

만약 공분산 행렬이 단위행렬(Identity matrix)이면 마하라노비스 거리는 유클리드 거리(Euclidean distance)가 된다. 또한 공분산 행렬이 직교(diagonal)일 때는 일반화된 유클리

드 거리(normalized Euclidean distance)가 된다.

$$d(\vec{x},\vec{y}) = \sqrt{\sum_{i=1}^{p} \frac{(x_i - y_i)^2}{\sigma_i^2}}$$

여기서 σ_i는 x_i의 표준편차이다.

일반적으로 큰 마하라노비스 거리 값을 갖는 관측치를 이상값이라고 한다. 그러나 실제에서 이 방법을 이용해서 이상 값을 관측하고자 할 때에는 masking effect와 swamping effect에 주의하여야 한다. masking effect는 마하라노비스 거리를 감소시켜서 이상 값인 관측치를 보통의 관측치로 감추어 버린다. 반면에 swamping effect는 마하라노비스 거리를 증가시켜 이상 값이 아닌 관측치를 이상 값으로 만든다. 이런 현상이 일어나는 이유는 μ와 $\sum$이 robust한 추정량이 아니기 때문이다. 따라서 masking effect와 swamping effect를 피하기 위해선 μ, $\sum$을 제외한 더 강력한 추정량을 사용하는 것이다. Campbell(1980), Stahel(1981), Donoho(1982), Hampel $et\ al$(1986), Rousseeuw and Leroy(1987)과 Rousseeuw and van Zomeren(1990)들에 의해서 몇 가지 추정량들이 제안되었다.

2. Hadi method

고전적인 이상치 탐색 방법은 데이터에 하나의 이상치가 있을 때 잘 찾아내는 방법이다. 그러나 하나 이상의 이상치가 존재 할 경우 이상치 찾는 검정력이 떨어진다. 그 이유는 바로 가림효과와 대세효과 때문이다. 그리하여 다변량 자료에서 여러 개의 이상치를 찾아내는 방법 중 효과적인 HADI(1992)의 방법을 제시한다.

HADI의 알고리즘은 다음과 같다. 우선 X라는 $n \times p$의 데이터 행렬이 있다고 가정하자. 여기에서 n은 관찰치 개수, p는 차원 수이다. 첫 번째로 각 n개의 관찰치들을 다음과 같은 로버스트한 거리를 이용하여 순서대로 오름차순 정리를 한다.

$$D_i(C_R, S_R) = \sqrt{(x_i - C_R)^T S_R^{-1} (x_i - C_R)}, \quad i = 1,2,\cdots,n$$

여기에서 C_R은 robust location, S_R은 covariance matrix estimator를 나타내는데, C_R과 S_R은 $D_i(C_M, S_M)$에서 기반한 추정량이다. 여기에서 C_M은 co-ordinatewise 중위수이며 S_M은 다음과 같다.

$$S_M = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - C_M)(x_i - C_M)^T$$

$D_i(C_M, S_M)$가지고 오름차순으로 관찰치를 정렬한 후 다음과 같은 가중함수를 정의한다.

$$v_i = \begin{cases} 1, & i\text{가 } (n+p+1)/2\text{의 정수 부분 보다 작을 때} \\ 0, & \text{그 밖의 경우} \end{cases}$$

그 다음 $C_R = C_V$, $S_R = S_V$ 로 놓으면 $D_i(C_R, S_R)$이 결정된다. 여기에서 C_V 와 S_V는 다음과 같다.

$$C_V = \frac{\sum_{i=1}^{n} v_i x_i}{\sum_{i=1}^{n} v_i}, \quad S_V = \frac{\sum_{i=1}^{n} v_i (x_i - C_V)(x_i - C_V)^T}{\sum_{i=1}^{n} v_i - 1}$$

이렇게 구해진 $D_i(C_R, S_R)$를 가지고 2개의 집단으로 나누게 된다. 하나의 집단은 $p+1$개, 또 다른 집단은 $n-p-1$개로 $p+1$개의 집단이 이상값이 없는 기본집단, $n-p-1$개는 비 기본집단이라고 한다. 두 번째는 기본집단의 중심에서 각각 관찰치의 값에서 상대적 거리를 구한다.

$$D_i(C_b, S_b) = \sqrt{(x_i - C_b)^T S_b^{-1} (x_i - C_b)}, \qquad i = 1,2,\cdots,n$$

여기에서 C_b는 기본집단의 평균,S_b는 기본집단의 공분산행렬이다.

세 번째는 두 번째에서 구한 식 $D_i(C_b, S_b)$에 의해 다시 n개를 오름차순으로 정리 한 후 기본집단은 $p+2$개, 비 기본집단은 $n-p-2$개로 기본집단의 크기를 증가시키며 집단을 나눈다. 이러한 과정이 계속 반복하다 다음과 같은 기준을 만족하게 되면 이 알고리즘은 멈추게 된다.

$$D_i(C_b, S_b) = \sqrt{(x_i - C_b)^T (c_b S_b)^{-1} (x_i - C_b)}, \qquad i = 1,2,\cdots,n$$

여기에서 c_b는 $= c_{npr} m_j / \chi^2_{pm0.5}$ 이며 $c_{npr} = \{1 + r/(n-p)\}^2$로 r은 마지막 기본집단의 관찰개수이다. 그래서 마지막으로 남는 비 기본집단이 그 자료에서의 이상값들이 된다.

HADI의 방법은 간단하면서도 계산이 쉽게 되면서 자동화가 가능하다. 더불어 여러 가지 소프트웨어에서 계산이 가능하기 때문에 매우 편리하다.

3. Local Outlier Factor method

다차원 자료에서 local outiler factor를 이용한 이상점을 찾는 방법은 LOF: Idenfifying Density-Based Local Outliers 논문에서 소개되었다.(Markus M.Breuning, 2000) 이 논문에서 소개된 방법은 자료에서 각 관찰값의 LOF(local outlker factor, 이후부터는 LOF로 표기함)값을 구하여 이상점 여부의 정도를 구별하는 것이다. LOF를 구하기 앞서 이 정의에

사용된 기본적인 정의를 살펴보면, p번째 관찰치의 k-distance는 임의의 양수 k에 대하여 관찰치 p의 k-distance는 k-distance(p)로 표기할 수 있다. k-distance(p)는 p와 관찰치 o 사이의 거리 d(p,o)로 정의 할 수 있다. k-distance가 주어졌을 때, 관찰치 p의 k-distance neighborhood는 p로부터의 모든 관찰값들의 거리는 k-distance보다 작다. MinPts는 밀도로 정의된 두 개의 모수에서 관찰치가 적은 모수를 의미한다. 이 정의들을 기초로 관찰치 p의 LOF는 다음과 같이 정의할 수 있다.

$$LOF_{MinPts}(p) = \frac{\sum_{o \in N_{MinPts}(p)} \frac{lrd_{MinPts}(o)}{lrd_{MinPts}(p)}}{|N_{MinPts}(p)|}$$

위의 식으로 구해진 LOF의 결과가 1에 가까울수록 이상점일 가능성이 높아진다. LOF의 값을 구하기 위해 사용된 lrd_{MinPts}는 관찰치 p의 도달 가능한 밀도를 나타내며 이는 다음과 같이 정의할 수 있다.

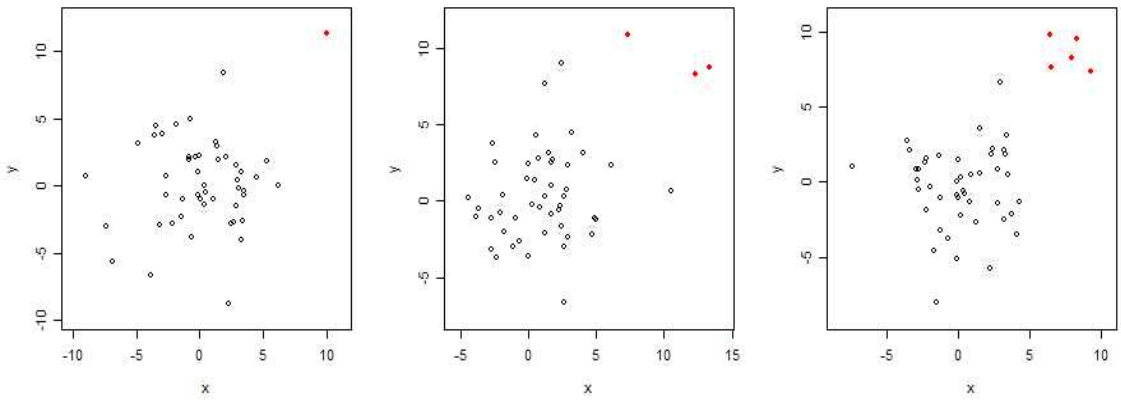
$$lrd_{MinPts}(p) = \frac{\sum_{o \in N_{MinPts}(p)} reach-dist_{MinPts}(p,o)}{|N_{MinPts}(p)|}$$

또한 관찰치 o로부터 관찰치 p까지의 도달 가능한 거리는 다음과 같이 정의할 수 있다.

$$reach-dist_k(p,o) = \max\{k-distance(o), d(p,o)\}$$

Ⅲ. 모의실험을 통한 이상점 탐색 방법들의 비교

모의실험을 통해서 위에서 살펴본 이상점 탐색 방법이 얼마나 잘 적용되는지를 알아보고자 한다. 모의실험은 이상점이 존재 하지 않는 그룹과 이상점이 존재 하는 그룹으로 나눠서 이상점을 탐색하는 정도를 알아보고자 한다. 또한 각 그룹의 총 자료의 수, 이상점의 개수와 이상점의 평균을 나눠서 각각의 방법의 장단점을 살펴보고자 한다. 자료의 개수는 $n = 50,100,1000$이며 각 집단별 이상점을 1, 3, 5개를 포함하였다. 자료는 이변량 정규분포에서 랜덤하게 추출 되었으며, 이상점들은 이상점이 아닌 자료로부터 평균이 떨어진 정도를 가지고 생성하였다. 아래의 그림은 각각 이상점이 주 관측치 보다 평균 10이 떨어진 경우로 각각 1, 3, 5개이고 50개의 자료를 그림으로 나타낸 것이다. 빨간 동그라미가 이상점으로 육안으로 쉽게 이상점을 구별해 낼 수 있다.



[그림 1] 모의실험을 위한 그룹 설정(n=50일 경우, 이상점 평균=10 이상점의 개수=1,3,5)

1. 이상점이 존재하지 않는 그룹에서의 이상점 탐색

이상점이 존재하지 않는 이변량 자료는 평균이 각각 0이고 이들의 공분산이 0.7인 이변량 정규분포 분포로부터 추출하고자 한다.

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \sim MN\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 10 & 0.7 \\ 0.7 & 9 \end{bmatrix}\right)$$

이상점이 존재 하지 않는 경우에는 모든 방법들이 이상점을 찾아내지 않는 것을 알 수 있다. 그리고 모든 방법들이 이상점이 아닌 값들에 대해서도 약간씩 이상점이라고 판단하는 경우가 있었다. 마할라노비스 거리의 경우 자료의 수가 커질수록 이상점이 아닌 경우를 이상점으로 잡는 오탐율이 높아지고 있다. 그러나 HADI 방법인 경우에는 자료의 수가 커질수록 오탐률이 현저히 작아짐을 알 수 있다.

<표 1> 이상점이 존재하지 않는 그룹에서의 이상점 탐색법비교

Method	data size = 50		data size =100		data size =1000	
	TRUE	FALSE	TRUE	FALSE	TRUE	FALSE
Mahalanobis	0.00000	0.00010	0.00100	0.00033	0.00300	0.00087
Hadi	0.00100	0.00108	0.00100	0.00053	0.00000	0.00005
LOF	0.00000	0.00065	0.00100	0.00082	0.00100	0.00034

2. 이상점이 존재하는 그룹에서의 이상점 탐색

이상점이 존재하는 그룹에서에서는 2개의 그룹으로 이루어진 이변량 자료를 생성하였다.

그룹1은 이상점이 아닌 자료들을 의미하며 그룹2는 이상점 그룹을 나타내며 이 분포의 평균을 각각 5, 10, 15로 증가시킴으로써 이상점을 만들어 하나의 데이터를 구성하였다. 이 때, 각각의 그룹들은 다음과 같은 분포를 가진다.

$$\text{그룹1} : \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \sim MN\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 10 & 0.7 \\ 0.7 & 9 \end{bmatrix}\right), \quad \text{그룹2} : \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \sim MN\left(\begin{bmatrix} 5 \\ 5 \end{bmatrix}, \begin{bmatrix} 10 & 0.7 \\ 0.7 & 9 \end{bmatrix}\right),$$

2.1 마할라노비스 거리

이상점 탐색방법으로 첫 번째인 마할라노비스 거리를 이용한 방법을 수행해 보았다.모의실험 결과를 자료의 수, 이상점의 평균 그리고 개수에 따라 살펴보고자 한다. 자료의 수가 커질수록 이상점을 더 잘 찾아냄을 알 수 있다. 그러나 오탐률 또한 높아지는 경향을 보였다. 이상점의 평균이 커질수록 정탐률이 점점 높아지고, 오탐률인 경우에는 점점 작아졌다. 이상점의 개수에 변화를 준 경우는 이상점의 개수가 많아질수록 정탐률이 많이 떨어지나 오탐률은 거의 비슷하다.

<표 2> 마할라노비스 거리를 이용한 이상점 탐색 결과

이상치개수	1 (n=50)		1 (n=100)		1 (n=1000)	
이상치평균	정탐률	오탐률	정탐률	오탐률	정탐률	오탐률
5	0.09100	0.00087	0.03400	0.00021	0.09100	0.00087
10	0.76800	0.00080	0.67100	0.00009	0.76800	0.00080
15	0.99900	0.00081	0.99200	0.00013	0.99900	0.00081
	5 (n=50)		5 (n=100)		5 (n=1000)	
이상치평균	정탐률	오탐률	정탐률	오탐률	정탐률	오탐률
5	0.00140	0.00009	0.01920	0.00014	0.08240	0.00081
10	0.00240	0.00000	0.1504	0.00015	0.73860	0.00071
15	0.00340	0.00007	0.37320	0.00005	0.99760	0.00049

2.2 Hadi Method

모의실험을 HADI 방법으로 한 결과 자료의 수가 커질수록 이상점을 더 잘 찾아내면, 오탐률도 작아짐을 알 수 있다. 그러나 자료의 수가 1000개일 경우 이상점 정탐률이 떨어짐을 발견할 수 있다. 이상점의 평균이 커질수록 정탐률이 높아지고, 오탐률도 증가한다. 특히 이상치의 평균이 10이 되면 평균 5를 가진 이상점들이 포함된 자료에 비해 정탐률이 급격하게 증가한다. 이상치의 평균이 15가 되면 90% 이상 이상점을 탐색해 내었다. Hadi method에 의한 이상점 탐색은 이상점의 개수가 많아질수록 정탐률이 떨어지고, 오탐률은 이상치의 개수가 변해도 거의 비슷하다.

<표 3> Hadi method를 이용한 이상점 탐색 결과

이상치개수	1 (n=50)		1 (n=100)		1 (n=1000)	
이상치평균	정탐률	오탐률	정탐률	오탐률	정탐률	오탐률
5	0.07300	0.00061	0.04500	0.00036	0.01800	0.00004
10	0.69600	0.00090	0.71700	0.00045	0.53200	0.00004
15	0.99400	0.00114	0.99400	0.00062	0.98800	0.00005
	5 (n=50)		5 (n=100)		5 (n=1000)	
이상치평균	정탐률	오탐률	정탐률	오탐률	정탐률	오탐률
5	0.02000	0.00053	0.02780	0.00031	0.01700	0.00005
10	0.35580	0.00089	0.51940	0.00039	0.50800	0.00004
15	0.98880	0.00111	0.99340	0.00052	0.99000	0.00006

2.3 Local Outlier Factor method

LOF 방법 또한 다른 방법과 마찬가지로 자료의 수가 커질수록 정탐률이 높다. 오탐률인 경우에는 자료의 개수가 증가함에 따라 일정한 패턴을 보이지 않았다. 이상점의 평균이 커지면 커질수록 정탐률이 점점 높아지고, 이상치의 평균이 15가 되면 LOF 방법도 90% 이상 이상점을 탐색해 내었다. 오탐률인 경우에는 증가하는 패턴을 보였다. 그리고 이상점의 개수가 많아지면 정탐률이 떨어진다. 오탐률의 경우 $n = 50, 100$에는 일정한 패턴을 보이지 않았으나, $n = 1000$ 인 경우에는 오탐률이 거의 비슷함을 알 수 있다.

<표 4> Local Outlier Factor method를 이용한 이상점 탐색 결과

이상치개수	1 (n=50)		1 (n=100)		1 (n=1000)	
이상치평균	정탐률	오탐률	정탐률	오탐률	정탐률	오탐률
5	0.04500	0.00057	0.05400	0.00075	0.05300	0.00035
10	0.59200	0.00027	0.68700	0.00068	0.62100	0.00035
15	0.98800	0.00045	0.99200	0.00089	0.99200	0.00040
	5 (n=50)		5 (n=100)		5 (n=1000)	
이상치평균	정탐률	오탐률	정탐률	오탐률	정탐률	오탐률
5	0.00840	0.00040	0.02260	0.00049	0.03500	0.00036
10	0.12320	0.00047	0.29460	0.00061	0.44140	0.00034
15	0.81500	0.00056	0.95100	0.00073	0.98680	0.00039

2.4 세 가지 방법의 비교

모의실험을 통해 이상점 탐색의 정,오탐률을 살펴본 결과 공통적으로 모든 방법에서 자료의 수가 커질수록 이상점을 더 잘 찾아냈다. 이상점의 평균이 커질수록 정탐률이 높아짐을 알 수 있다. 또한 이상점의 개수가 증가 할 때는 정탐률이 떨어짐을 도출 할 수 있다. 모의실험의 결과를 자료의 개수, 이상점의 평균과 개수의 관점에서 세가지 방법들을 비교해보고자 한다.

첫 번째로 자료의 개수에 따른 결과 비교이다. 같은 자료의 개수에서 비교하였을 때 마할라노비스 거리보다 HADI방법과 LOF의 방법이 더 우수하게 이상점을 잘 찾아냈으며. HADI 방법과 LOF 방법 중에는 HADI방법의 정탐률이 높았다.

두 번째로 이상점 평균 거리의 차이의 관점에서는$n = 50, 100$인 경우에는 마할라노비스 거리보다 HADI 방법과 LOF방법이 이상점의 평균 거리가 멀어질수록 정탐률이 좋았다. 그러나 오탐률도 높아짐을 알 수 있다. 반면 HADI 방법은 정탐률도 좋을 뿐 아니라 $n = 1000$에서의 오탐률을 살펴보면 다른 방법에 비해 현저히 낮은 결과를 얻었다.

세 번째 이상점 개수의 변화에 따른 세 방법의 비교의 결과를 살펴보면 마할라노비스 방법은 $n = 50, 100$ 일 때, HADI방법과 LOF 방법보다 이상점이 1개 일 때보다 5개 일 때 정탐률이 많이 떨어진다. 즉 마할라노비스 방법에서 이상치가 여러 개 있을 때 잘 찾아내지 못함을 알 수 있다. HADI 방법과 LOF 방법은 이상점의 개수가 증가하면 조금씩 정탐률이 떨어지지만 큰 차이를 보이지 않고 있다. 즉 이 세 가지 방법 중에서 자료의 수가 크고 이상점이 많은 경우에는 HADI 방법이 적합함을 알 수 있다.

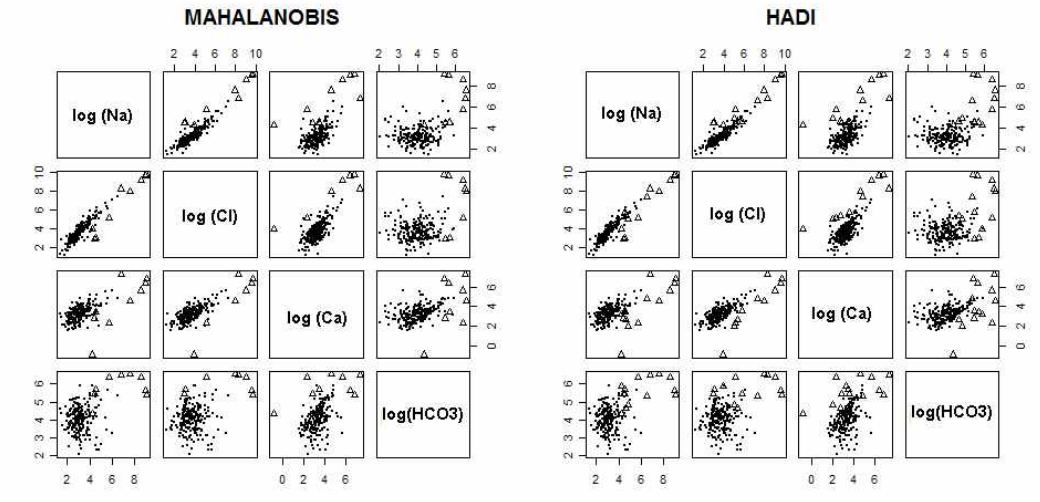
Ⅳ. 실제 사례 데이터를 이용한 이상점 탐색방법들의 비교

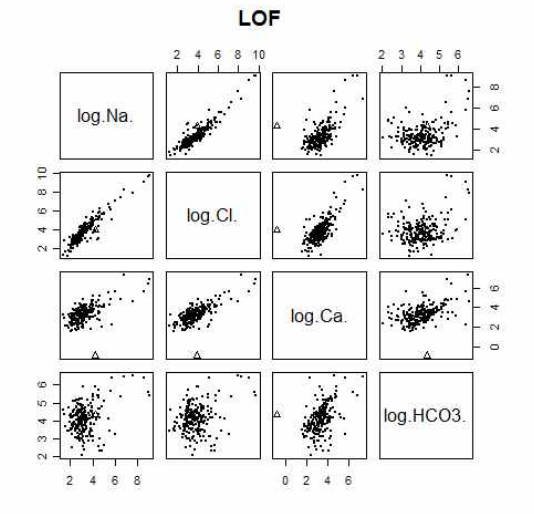
서해는 염류화의 많은 염류화의 요인들을 갖고 있다. 서해의 해안 모양은 곡선의 형태이며 특히 남서쪽은 전형적인 리아스식 해안의 형태를 하고 있다. 또한 6m 이상의 큰 조수간만의 차를 갖고 있기 때문에 바닷물이 강 상류까지 흘러들어오기도 한다. 그리고 열악한 하수처리 시설과 한강과 금강, 영산강과 같은 주요 강들로부터 많은 양의 침전물과 오염물질, 충적토들이 들어오게 된다. 이런 여러 가지 이유들로 인해 서해는 염류화의 위험이 매우 높다.

서해안 지역에 있는 지하수들의 염류화(salinization)정도를 추정하기 위해서 비가 잘 내리지 않는 2000년 3월과 5월 사이에 그리고 2000년 11월, 2001년 1월과 2월에 서해안의 해안으로부터 10km 이내에 있는 얕은 우물(깊이가 100m이하인)들을 표본으로 추출하였다. 샘플로 뽑힌 지하수들은 측정도구들을 이용하여 PH와 전기전 전도도(electrical conductivity)을 측정한 후 얇은 섬유소 막에 필터링 한 후에 폴리에틸렌 병에 넣어 보관하였다. 288개의 자료는 Na, Ca, Cl, HCO_3, NO_3 총 5개의 변수가 있으나 NO_3 를 제외한 4개 원소에 관해서만 분석 시행하였다. 또한 지질학적 연구의 자료들은 로그 변환되어 사용하는 것이 좋다는 기존 연구를 참조하여 로그 변환된 값들을 사용하였다.

실제 사례 데이터에 마할라노비스방법을 적용시켜 이상점을 찾아낸 경우는 9가지의 이상점을 탐색하였다. 이상점으로 찾아낸 관측치는 9, 23, 29, 52, 53, 116, 125, 147, 213 번이다. HADI 방법으로는 총 14개의 이상점으로 8, 9, 23, 29, 38, 52, 53, 116, 125, 138, 147, 208, 212, 213번으로 판명 되었다. LOF 방법으로는 1개의 이상점으로 213 번째의 관측치가 이상점으로 발견되었다. 이렇게 선택된 이상점들을 살펴보면 Na, Ca, Cl, HCO_3의 수치가 다른 관측치에 비해 매우 높은 수치를 가지고 있음을 알 수 있다. Na인 경우 보통 관측치에 비해 평균 52.7224배, Cl은 47.847배, Ca는 7.124배 HCO_3는 4.708배로 높은 수치를 가졌다. 이는 분명 이상점이라고 볼 수 있다. 즉, 이러한 이상점들은 서해안 지역에 있는 지하수들의 염류화정도가 보통 다른 관측치에 비해 높은 수치를 가지고 있는 것이다. 그러므로 이들을 이상점으로 분류하는 것은 합당하다고 볼 수 있다.

이상점 탐색 방법 간 비교를 해보면 마할라노비스방법보다 HADI 방법이 더 정교하게 이상점을 탐색하였다. LOF 방법인 경우에는 여러 이상점이 있는 데도 불구하고 이상점을 잘 찾는데 못했다.





[그림 2] 실제 데이터를 적용한 이상점 탐색결과

Ⅴ. 결 론

본 연구에서는 다변량자료에서의 이상점 탐색에 관한 세 가지 방법들에 대해 모의실험과 실제 자료에 적용시켜 각각의 이상점 탐색의 정,오탐률을 살펴보았다. 모의실험의 결과 마할라노비스 거리, HADI방법과 LOF 방법는 공통적으로 자료의 수가 커질수록 이상점을 더 잘 찾아낼 수 있다. 이상점의 평균이 커지면 커질수록 정탐률이 점점 높아짐을 알 수 있다. 또한 이상점의 개수가 증가 할 시에는 정탐률이 떨어짐을 도출 할 수 있다. 서해안 지역의 염류화에 대한 실제 자료에 대해 이상점 탐색 방법을 적용한 결과

마할라노비스방법보다 HADI 방법이 더 정교하게 이상점을 탐색하였다. 반면 LOF 방법은 여러 이상점이 있는 데도 불구하고 이상점을 잘 찾는데 못했다. 위의 결과로 보아 이상점 탐색방법에서는 HADI 방법이 우수함을 알 수 있다.

본 연구에서는 다변량 자료에서의 이상점 탐색에 관한 세 가지 방법만을 다루었고, 모의실험에서는 이상치의 개수와 평균을 좀 더 세분화 시키지 못한 한계점을 가지고 있다. LOF 방법에서 상,하한 값 설정에서 기본값을 사용하지 않고, 그 값을 도출하여 이상점 탐색을 해보는 것은 좀 더 의미있는 연구가 될것으로 생각되어진다.

참고문헌

[1] Hadi, A. S. (1992), "Identifying Multiple Outliers in Multivariate Data," Journal of the Royal Statistical Society, Series (B), 54, 761-771.

[2] Hadi, A. S. (1994), "A Modification of a Method for the Detection of Outliers in Multivariate Samples," Journal of the Royal Statistical Society, Series (B), 56, 393-396.

[3] *Osmar Zaïane*(1999,2007) "Many names for Outlier Detection". Principles of Knowledge Discovery data.

[4] E.S. Gillespie(1993) "An application of multivariate outlier detection in assessing family characteristics for bank advertisements". The Statistician 42, 231-235

[5] Edgar Acuna and Caroline Rodriguez "A Meta analysis study of outlier detection methods m classification".

[6] Seh-Chang Park et al.(2005) "Regional hydrochemical study on salinization of coastal aquifers, western coastal area of South Korea" Journal of Hydrology 313 182-194

[7] Markus M. Breunia, Hans-Peter Krieael, Ravmond T.Na, *JörgSander* (2000) "LOF: Identifying Density-Based Local Outliers"

[8] Pang-Ning Tan ,Michael Steinbach, Vipin Kumar ," Introduction to Data Mining"

[투고일: 2009. 2. 2] [심사일: 2009. 2.10] [게재확정일: 2009. 2.15]

Annals of Statistical Research
Vol. 5. No. 1, 2009, pp.33~44

Comparison of Allocation Methods for Enterprise selection

Eun-Joo Kwak[1], In-Ki Kim[2], Hea-Jung Kim[3]

Abstract

The case where the size of population stratum differs extremely from stratified random sampling compared Neyman allocation where is a sample allocation method could be used, Power allocation and Proportional allocation.

The data which uses in simulation stratum with 5 and the sample used RMT85 and REV84 variables where the correlation is high with MU284 population allocated. In the Study not only the whole character which the sample has in the hazard Neyman allocation which grasps the character which stratum has power differently with sample whole or it knows stratum sample variable, sample error and coefficient of variation and the paper it does.

Key words : Neyman allocation, Power allocation, Proportional allocation, MU284 population

1) Doctoral Candidate for Statistics, Department of Statistics, Dongguk University, Seoul 100-715, Korea. E-mail: hwaryun@dongguk.edu
2) Doctoral Candidate for Statistics, Department of Statistics, Dongguk University, Seoul 100-715, Korea. E-mail: inkey@dongguk.edu
3) Associate Professor, Department of Statistics, Dongguk University, Seoul 100-715, Korea. E-mail: kim3hj@dongguk.edu

Ⅰ. 서 론

사업체 조사에 흔히 이용되는 모집단은 통계청의 사업체기초통계조사결과로 매년 전국 및 시·도별 산업별 사업체명, 종사자수, 총매출액 등 10개 항목을 조사하는 조사(전수)통계이다. [그림 1]과 [그림 2]는 2007년 사업체기초통계조사 자료와 그래프로 행정구역별 산업별 사업체구분별로 층화한 후 각 층의 사업체수, 종사자수를 제공하고 있다.
(www.kosis.kr)


위치 : 각 시도 > 사업체기초통계조사

통계설명자료 | 도움말

분류 설정 | 조회범위 상세설정하기 | 분류와 항목의 조회범위를 재설정하여 원하는 상세자료를 볼수 있습니다.

조회 기간 | 년 | 2007 ~ 2007 | 수록기간 (2007 ~ 2007)

소수점 | 한글보기 | 계층별 컬럼 구분

자료다시 불러오기

자료받기 | 자료분석 | 찾기 | 표 구성 | 기본 통계표

표측 | 행정구역별 | 산업별 | 사업체구… 표두 | 년 | 항목 숨김

시도·산업·사업체구분별 사업체수, 종사자수('07, 9차 개정)

행정구역별	산업별	사업체구분별	2007	
			사업체수 (개)	종사자수 (명)
전 국	전산업	계	3,262,925	15,928,985
		단독사업체	3,117,390	11,830,830
		공장,지사,영업소	117,956	2,469,448
		본사,본점	27,579	1,628,707
	농업, 임업 및 어업 (01…	계	2,274	34,103
		단독사업체	1,737	22,590
		공장,지사,영업소	473	8,731
		본사,본점	64	2,782
	광업 (05 ~ 08)	계	1,790	17,937
		단독사업체	1,585	9,944
		공장,지사,영업소	129	5,731
		본사,본점	76	2,262
	제조업 (10 ~ 33)	계	334,227	3,396,087
		단독사업체	314,040	2,056,129
		공장,지사,영업소	11,167	705,141
		본사,본점	9,020	634,817
	전기, 가스, 증기 및 수…	계	1,570	68,930

출처 : 통계청 경제통계국 산업통계과 (☎042-481-2145) +더보기

서비스 기관 : 통계청

그림 1. 전국사업체기초통계조사(2007년)

그림 2. 전국사업체기초통계조사 그래프(2007년)

표본설계를 위해 사업체기초통계조사결과를 산업별로 층화한 후 각 산업에 할당된 표본을 종사자 규모 층에 할당하게 되는데 이때 종사자 규모별 부모집단의 크기는 $N_1 > N_2 > \cdots > N_h > \cdots > N_L$의 관계를 보이는 것이 일반적이다. 그러나 이와는 반대로 종사자 수에 대한 분산은 $S_1^2 < S_2^2 < \cdots < S_h^2 < \cdots < S_L^2$ 의 관계를 보이는 것이 일반적이다. 즉 부모집단 크기와 분산의 관계는 역의 관계를 보이고 있지만 그 양적 크기는 $N_h < S_h^2$ 의 관계를 보이게 되어 부모집단(subpopulation)의 크기보다 층의 분산이 상대적으로 매우 크다는 것을 알 수 있다. 이와 같은 관계는 종사자수 뿐만 아니라, 매출액 등에서도 같은 현상을 보이고 있다.

사업체 조사 분야에서 주로 사용되고 있는 표본할당방법은 네이만할당(Neyman allocation)으로 부모집단 크기와 분산의 관계는 역의 관계를 보이는 특성을 갖는 사업체조사에서 네이만할당으로 각 층에 표본을 할당 하면 각 층에 할당된 표본의 크기는 $n_1 > n_2 > \ldots > n_h > \ldots > n_L$ 의 관계를 갖기 때문에 사업체조사에서 특정산업의 종사자 수가 1~4인, 5~9인 등의 종사자규모가 작은 층1(N_1)에 대부분의 사업체가 할당되어 종사자 수가 300인 이상, 500인 이상 등의 종사자규모가 큰 층L(N_L)에 극소수의 사업체가 할당되든가 또는 전혀 할당되지 않는 경우가 발생한다. 이 경우, 극소수 추출된 사업체가 무응답을 제공하게 되면 표본의 대표성 문제를 더욱 심각하게 만든다.

대부분의 경우 특성치의 비율이 규모가 큰 층에 적절한 크기의 표본이 추출되어야 신뢰성(reliability)을 확보 할 수 있다.

네이만할당(Neyman allocation)은 부모집단의 크기에 비례하여 할당하므로 층별 분산의 크기가 유사한 경우, 부모집단의 크기가 작은 층에 매우 적은 표본사업체가 할당되어 해당 층의 정도(precision)를 감소시키는 결과를 초래한다. 반면에 종사자규모가 작은 층(부모집단의 크기가 큰 층)에는 필요이상의 표본사업체가 할당되어 층별 산업 추정치($\overline{y}_{st,\text{산업}}$)의 비용 대비 정도를 떨어뜨리는 결과를 초래하게 된다. 즉 표본크기가 필요이상으로 크면 추정의 정도는 높아지나 과다한 비용이 요구되며, 반대로 표본의 크기가 너무 작으면 비용은 감소하나 추정의 정도가 작아진다.

본 연구에서는 층의 규모가 작은 층에 표본이 할당되지 않는 문제를 해결하기 위해 규모가 매우 작은 층을 전수층으로 하고, 네이만할당, 승수할당(Power allocation), 비례할당(Proportional allocation) 세 가지 방법으로 각 층에 표본사업체를 할당한 후 표본분산, 표본오차, 변동계수 추정 결과를 비교해 보고자 한다.

Ⅱ. 본 론

2.1 자료설명

표본할당법 비교를 위해 이용된 조사모집단은 "MU284 population" 자료(출처 : "Model Assisted Survey Sampling" by Särndal, C.-E., Swensson, B., and Wretman, J. (1992)) 이다. 이 자료는 스웨덴의 284개 지자체를 대상으로 1985년 인구, 1975년 인구, 자치 시의회 보수당 의석 수, 1984년 종사자 수 등 10개 변수로 구성되어있다(표 1). 이 중 상관관계가 높은 '1985년 도시 과세 수익(RMT85)'와 '1984년 부동산 평가액(REV84)' 변수를 이용하였다. 자료 분석은 SAS 9.1 버전을 사용하였다.

[표 1] MU284 Population 자료

변수	설명
P85	1985년 인구(천명)
P75	1975년 인구(천명)
RMT85*	1985년 도시 과세 수익(수 백만 크로네)
CS82	자치 시의회 보수당 의석 수
SS82	자치 시의회 사회민주당 의석 수
S82	자치 시의회 전체 의석 수
ME84	1984년 행정직원의 수
REV84*	1984년 부동산 평가액(수 백만 크로네)
REG	지리적인 지역 표시기
CL	클러스터 표시등 (클러스터 인근 집합)

주) * : 표본할당방법 비교에 이용한 변수

2.2 표본할당법

층화임의추출에서 모집단 층의 크기가 극단적으로 상이한 경우 사용될 수 있는 표본할당법으로 네이만할당법(Neyman allocation), 비례할당법(Proportional allocation), 승수할당법(Power allocation)이 있다.

네이만할당법은 각 층의 분산 혹은 표준편차를 알고 있어야 사용 가능한 방법으로 각 층의 크기와 층별 변동의 정도를 동시에 고려한 표본배정 방법이다. 또한 각 층별 조사비용은 별 차이가 없고, 층별 변동의 정도가 많이 나는 경우에 적당하다. 표본크기를 결정하는 식 (1.1)과 같다.

$$n_h = n \cdot \frac{N_h S_h}{\sum_{h=1}^{L} N_h S_h} \qquad (1.1)$$

여기서 N_h : h층의 부모집단 크기, $h = 1, \cdots, L$

S_h : h층의 표본분산

n_h : h층의 표본의 크기

L : 층의 수

n : 표본크기

비례할당은 기본적으로 각 층의 분산이 같고 각 층의 부모집단 크기가 다를 때 이용하는 방법으로 모집단에 대한 각 층의 크기에 비례하여 배정하는 방법으로 시행하기 쉽고 집계하는데 간편하며 크기가 작은 층이 크기가 큰 층보다 작다는 가정을 전제로 한다. 표본 크기를 결정하는 식(1.2)와 같다.

$$n_h = n \cdot \frac{N_h}{N} \tag{1.2}$$

여기서 N : 모집단의 크기($\sum_{h=1}^{L} N_h$)

N_h: h층의 부모집단 크기, $h = 1, \cdots, L$

n_h: h층의 표본의 크기

n : 표본크기

승수할당은 네이만할당을 응용한 것으로 부모집단의 크기가 지나치게 크거나, 부모집단의 분산이 지나치게 클 때 특정 층에 지나치게 많은 표본이 할당되는 것을 방지하면서 네이만할당의 장점과 비례할당의 장점을 동시에 갖춘 방법으로 식(1.3)과 같다. 만약 $p=1$ 이면 네이만할당법과 동일하다.

$$n_h = n \cdot \frac{(N_h S_h)^p}{\sum_{h=1}^{L} (N_h S_h)^p} \quad , \ 0 \leqq p \leqq 1 \tag{1.3}$$

여기서 N_h : h층의 부모집단 크기, $h = 1, \cdots, L$

S_h : h층의 표본분산이다.

n_h : h층의 표본의 크기

L : 층의 수

n : 표본크기

p : 승수

2.3 표본크기

층별 모집단 크기와 표본크기 추출은 자료 'MU284 population'에서 '1984년 부동산 평가액(REV84)'을 이용하여 5개 층에 모집단 크기를 각각 $N_1=87$, $N_2=82$, $N_3=65$, $N_4=45$, $N_5=5$ 로 하고, 표본크기는 네미만할당법, 승수할당법, 비례할당법으로 표본크기를 추출하였다(표 2).

[표 2]에서 승수할당법으로 추정한 표본크기는 승수(power) 값에 따라 차이를 보이고 있다. 승수할당법에 의한 승수를 0.1≤승수(p)≤0.3 일 때, 승수(p)=0.4, 승수(p)=0.5 일 때, 승수(p)=0.6, 승수(p)=0.7 일 때, 승수(p)=0.8, 승수(p)=0.9 일 때, 4가지로 구분하였는데 이것은 승수를 0 ≤ 승수(p) ≤1 범위에서 0.1씩 달리 주었을 때 추출된 표본크기가 동일하였기 때문이다(표 2). 여기서 층5는 전수층으로 층에 속한 모든 개체가 표본으로 추출된다. 이는 층의 모집단 크기가 다른 층과 비교 했을 때 상대적으로 작은 층을 전수층으로 하여 정도가 높은 추정치를 얻기 위해서 이다.

[표 2] 층별 모집단크기 및 추출 표본크기

	할당법	층(h)				
		1	2	3	4	5†
부모집단크기 (N_h)		87	82	65	45	5
표본크기(n_h)	네이만할당	2	2	3	7	5
	승수할당 (0.1≤p≤0.3)	3	3	4	4	5
	승수할당 (p=0.4, p=0.5)	3	3	3	5	5
	승수할당 (p=0.6, p=0.7)	2	3	3	6	5
	승수할당 (p=0.8, p=0.9)	2	2	3	7	5
	비례할당	4	4	3	3	5

주) †는 전수층, p : power

2.4 시뮬레이션결과

시뮬레이션 결과는 '1985년 도시 과세 수익(RMT85)' 자료를 이용하여 앞서 설명한 세 가지 방법으로 표본분산, 표본오차, 변동계수를 추정하여 보았다(표 3 ~ 표 8).

[표 3]은 네이만할당법으로 추정한 층별 표본 총분산, 표본오차, 변동계수 추정량들을 정리한 표이다. 전체 변동계수 값은 '16.76', 각 층별 변동계수 값은 층4(5.92)< 층3(8.52) < 층2 (12.33)< 층1(13.21)순으로 낮았다.

[표 4]는 승수할당법으로 0.1 ≤승수 ≤0.3 일 때 층별 표본 총분산, 표본오차, 변동계수 추정량들을 정리한 표이다. 전체 변동계수 값은 '18.99', 각 층별 변동계수 값은 층3 (7.38)< 층4 (8.17)< 층2(10.05) < 층1(10.76) 순으로 낮았다.

[표 5]는 승수=0.4, 승수=0.5 일 때 층별 표본 총분산, 표본오차, 변동계수 추정량들을 정리한 표이다. 전체 변동계수 값은 '16.76', 각 층별 변동계수 값은 층4 (7.37)< 층 3(8.52) < 층2(10.05) < 층1(10.76) 순으로 낮았다.

[표 6]은 승수=0.6, 승수=0.7 일 때 층별 표본 총분산, 표본오차, 변동계수 추정량들을 정리한 표이다. 전체 변동계수 값은 '16.45', 각 층별 변동계수 값은 층4 (6.54)< 층3(8.52) < 층2(10.05) < 층1(13.21) 순으로 낮았다.

[표 7]은 승수=0.8, 승수=0.9 일 때 층별 표본 총분산, 표본오차, 변동계수 추정량들을 정리한 표이다. 전체 변동계수 값은 '21.54', 각 층별 변동계수 값은 층4 (5.92)< 층3(8.52) < 층2(10.05) < 층1(13.21) 순으로 낮았다.

[표 3] 층별 추정량

네이만할당			
층	총분산	표본오차	변동계수
1	60.49	15.24	13.21
2	173.34	25.81	12.33
3	329.38	35.57	8.52
4	919.14	59.42	5.92
5†	-	-	-
전체	1482.35	75.46	16.76

주) †는 전수층

[표 4] 층별 추정량

승수할당 (0.1≤p≤0.3)			
층	총분산	표본오차	변동계수
1	39.79	12.36	10.76
2	114.77	21.00	10.05
3	247.59	30.84	7.38
4	1763.12	82.30	8.17
5†	-	-	-
전체	2165.28	91.20	18.99

주) p : power, †는 전수층

[표 5] 층별 추정량

승수할당 (p=0.4, p=0.5)			
층	총분산	표본오차	변동계수
1	39.79	12.36	10.76
2	114.77	21.00	10.05
3	329.38	35.57	8.52
4	1420.12	73.86	7.37
5†	-	-	-
전체	1904.06	85.53	17.85

주) p : power, †는 전수층

[표 6] 층별 추정량

승수할당 (p=0.6, p=0.7)			
층	총분산	표본오차	변동계수
1	60.49	15.24	13.21
2	114.77	21.00	10.05
3	329.38	35.57	8.52
4	1108.14	65.25	6.54
5†	-	-	-
전체	1612.78	78.71	16.45

주) p : power, †는 전수층

[표 7] 층별 추정량

승수할당 (p=0.8, p=0.9)			
층	총분산	표본오차	변동계수
1	60.49	15.24	13.21
2	173.34	25.81	12.33
3	329.38	35.57	8.52
4	919.14	59.42	5.92
5†	-	-	-
전체	1482.35	75.46	16.76

주) p : power, †는 전수층

[표 8] 층별 추정량

비례할당			
층	총분산	표본오차	변동계수
1	30.83	10.88	9.45
2	86.29	18.21	8.62
3	329.38	35.57	8.52
4	2356.46	95.15	9.51
5†	-	-	-
전체	2802.95	103.77	21.64

주) p : power, †는 전수층

[표 8]은 비례할당법으로 추정한 층별 분산추정치, 표본오차 그리고 변동계수 추정량들을 정리한 표이다. 전체 변동계수 값은 '21.64', 각 층별 변동계수 값은 층3 (8.52)< 층2 (8.62)< 층1 (9.45)< 층4(9.51) 순으로 낮았다. 즉 층3의 표본의 추정 변동계수 값이 '8.52'으로 가장 작았다.

[표 9] 네이만할당을 기준으로 승수할당 변동계수 비교

	변동계수 감소	변동계수 증가
	승수	승수
층1	0.1, 0.2, 0.3, 0.4, 0.5	
층2	0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7	
층3	0.1, 0.2, 0.3	
층4		0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7
층5	-	-
전체	0.6, 0.7	0.1, 0.2, 0.3, 0.4, 0.5

[표 9]는 네이만할당법으로 구한 변동계수 값을 기준으로 승수할당법으로 구한 변동계수 값이 작아진 승수들을 정리한 표이다. 결과는 전체 변동계수 값을 살펴보면 승수=0.6, 승수=0.7 일 때, 층별로 변동계수 값을 살펴보면 층4를 제외한 층1은 0.1 ≤승수≤0.5, 층2는 0.1 ≤승수≤0.7, 층3은 0.1 ≤승수≤0.3 일 때 네이만할당법 보다 승수할당법을 이용하였을 때 변동계수 값이 작았다.

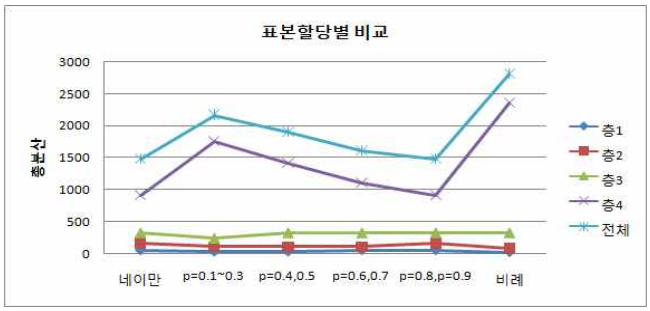


그림 4. 표본할당 방법에 따른 총분산

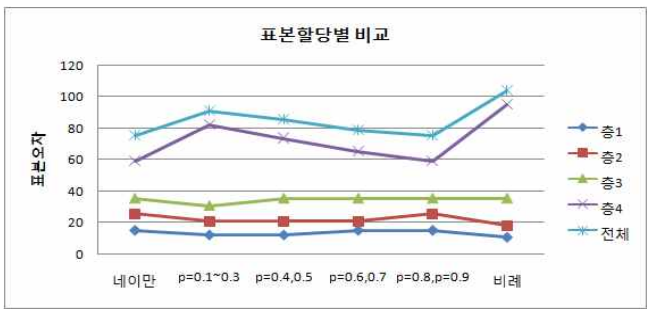


그림 5. 표본할당 방법에 따른 표본오차

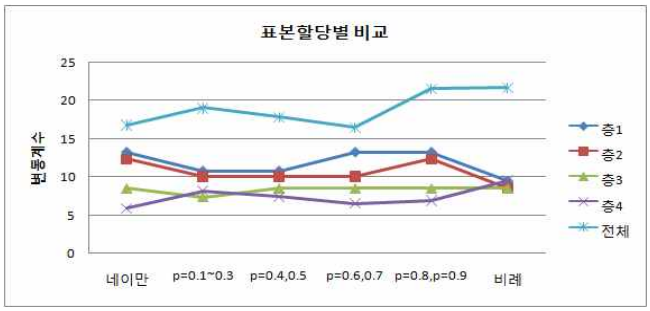


그림 6. 표본할당 방법에 따른 변동계수 비교

Ⅲ. 결 론

현재 사업체조사에서 사용되는 층화임의추출방법인 네이만할당법, 비례할당법, 승수할당법을 이용하여 모집단을 5개 층으로 층화하고, 모집단 수가 작은 층을 전수층으로 하여 표본크기를 추출하였다. 추출된 표본으로 각 층의 총분산, 표본오차 그리고 변동계수 값을 비교하였다. 네이만할당법으로 추출된 표본의 전체 변동계수가 '16.76', 비례할당법으로 추출된 표본의 전체 변동계수가 '21.64', 승수할당법으로 추출된 표본에서 $0.1 \leq p \leq 0.3$ 일 때 전체 변동계수가 '18.99' 이고, 승수(p)=0.4, 승수(p)=0.5 일 때 전체 변동계수 가 '17.85' 이고, 승수(p)=0.6, 승수(p)=0.7 일 때 전체 변동계수가 '16.45' 이고, 승수(p)=0.8, 승수(p)=0.9 일 때 전체 변동계수가 '21.54' 이였다. 층별로 변동계수 값을 살펴보면 층4를 제외한 층1은 0.1 ≤승수≤0.5, 층2는 0.1 ≤승수≤0.7, 층3은 0.1 ≤승수≤0.3 일 때 네이만할당법 보다 승수할당법을 이용하였을 때 변동계수 값이 작았다.

사업체조사에서 세 가지 표본할당방법 중 승수할당법은 주어진 비용하에서 변동계수를 최소화시키는 것으로 나타났는데 이 중 전체 변동계수가 가장 작은 승수할당은 승수(p)=0.6, 승수(p)=0.7 일 때 인 것으로 나타났다.

본 논문에서 사업체 조사 분야에서 분산과 비용을 적절히 고려한 표본설계 방법으로 네이만할당 식에 승수를 달리 줌으로써 전체 모집단에 대한 정도 높은 즉 분산을 작게 하고 표본오차, 변동계수 값들을 작게 하는 효과가 있는 것으로 볼 수 있을 것이다. 또한 승수할당법은 층별로 분산, 표본오차, 변동계수를 작게 하는 것을 알 수 있었다. 이 결과로 부터 부모집단의 오차를 줄이는데 활용해 봄으로써 부모집단에 대한 정확성과 이해성을 높일 수 있는 효과가 있을 것으로 기대된다. 또한 표본조사에서 드는 비용과 시간을 줄이는데 효과를 볼 수 있을 것이다.

참고문헌

[1]. Cochran, W.G. (1977). Sampling Techniques, 3rd edition. New York: John Wiley and Sons, Inc.

[2]. Cox, B.G., Binder, D.A., Chinnappa, B.N., Christianson, A., Colledge, M.J., and Kott, P.S. (1995). Business Survey Methods. New York: John Wiley and Sons, Inc.

[3]. Groves, R.M. (1989). Survey Errors and Survey Costs. New York: John Wiley and Sons, Inc.

[4]. Hansen, M.H Hurwitz, W.N.; Madow. W.G.; (1953); Sample Survey Methods and Theory; John Wiley and Theory. Volume I: Methods and Applications. Volume II: Theory New York: John Wiley and Sons, Inc.

[5]. Jessen, R. J. (1978). Statistical Survey Techniques. New York: John Wiley and Sons, Inc.

[6]. Kish, L. (1965). Survery Sampling. New York: John Wiley and Sons, Inc.

[7]. Lohr, S.L(1999). Sampling: Design and Analysis. Pacific Grove: Brooks/Cole Publishing company.

[8]. Särndal, C.-E., Swensson, B., and Wretman, J. (1992), Model Assisted Survey Sampling, Springer Verlag, New York

[9]. Scheaffer, R.L., Mendenhall, W., and Ott, L. (1996). Elementary Survey Sampling, 6th edition. Boston: PWS Kent.

[10] 통계청 국가 통계 포털 사이트 http://www.kosis.kr/

[투고일: 2009. 2. 2] [심사일: 2009. 2.10] [게재확정일: 2009. 2.15]

Annals of Statistical Research
Vol. 5. No. 1, 2009, pp.45~53

On residuals analysis in linear mixed models for longitudinal data

Hong-Yup Ahn[1], Min-Su Kang[2]

Abstract

We widely use linear models as powerful tools to analyze data in various fields. This is because that linear models enable us not only to demonstrate the relation of cause and effect related to phenomena but also to interpret models readily. Many diagnostics have also been developed for linear models, such as residuals analysis, leverage, and Cook's distance. When assumptions on residuals of linear models do not be satisfied, we would consider another models fitted for data. In the study, we especially focus on the normality of residuals in longitudinal data cases. In particular, we are interested in normality assumption of linear mixed models which could be used to reflect the within-group correlation for longitudinal data.

Keywords : linear mixed models, longitudinal data, normality, residuals analysis.

1) (Corresponding author) Assistant Professor, Department of Statistics, Dongguk University, Seoul, 100-715, Korea. E-mail: ahn@dongguk.edu
2) Master's course, Department of Statistics, Dongguk University, Seoul 100-715, Korea

1. 서 론

선형 모형은 통계학에서 많이 사용되는 모형 중의 하나로 사회학, 금융, 경영, 의학 등 폭넓게 사용되고 있다. 선형 모형을 통해 종속변수에 대한 독립변수의 영향을 확인할 수 있으며 이러한 관계를 규명함으로써 특정 결과나 현상의 인과관계를 밝힐 수 있다. 선형 모형의 특징 중의 하나로 모형의 해석이 간단하다는 점을 들 수 있다. 이는 복잡한 자료로부터 원하는 정보를 구하는데 있어 직관적이고 단순한 모형을 사용하여 문제 해결을 도모한다는 점에서 매우 매력적인 모형이라고 할 수 있겠다.

선형 모형의 대표적인 활용예로써 선형회귀 모형, ANOVA 모형, ANCOVA 모형 등을 생각해 볼 수 있다. 각 모형들은 단일 형태로

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{1}$$

으로 표현할 수 있다. 여기서, $\mathbf{y}$는 $n \times 1$의 종속변수 벡터, $\mathbf{X}$는 $n \times p$ 독립변수 행렬, $\boldsymbol{\beta}$는 $p \times 1$ 회귀계수 벡터, 그리고 $\boldsymbol{\epsilon}$은 $n \times 1$ 오차항 벡터를 나타낸다. $\mathbf{y}$는 수치형 자료이고 일반적으로 $\boldsymbol{\epsilon}$의 정규성에 의해 정규분포를 따르는 것으로 가정한다. $\boldsymbol{\epsilon}$은 n개의 오차항으로 이루어져 있는데 이들은 서로 독립이고 기대값과 분산이 각각 0과 σ^2인 정규분포를 따른다고 가정한다.

$\mathbf{X}$가 수치형 자료들을 나타내는 경우 모형 (1)은 선형회귀 모형이라 부른다. 이 경우 종속변수와 각각의 독립변수 간의 상관관계를 대응되는 회귀계수로 규명할 수 있다. 특정 종속변수에 상관관계가 있을 것으로 생각되는 여러 수치형 자료들 중에 통계적으로 유의한 상관관계를 갖는 변수를 찾는데 유용하게 사용될 수 있다. 뿐만 아니라 종속변수와 독립변수 간의 기울기를 구하여 향후 독립변수의 변화에 따른 종속변수의 변화를 구체적으로 제시할 수 있다 (Kutner *et al.,* 2005).

$\mathbf{X}$가 범주형 자료들로 이루어진 경우에는 모형 (1)을 ANOVA 모형이라 하고 $\mathbf{X}$는 범주형 변수의 범주(수준) 갯수에 따라 생성되는 지시변수들로 이루어져 있다. 수치형 독립변수가 $\mathbf{X}$의 각 열을 구성하는 회귀모형과 달리 여러 개의 열 즉 지시변수들이 하나의 범주형 자료에 연관되어 있다. 이 때문에 각각의 지시변수들의 유의성을 확인하는 것보다 연관된 지시변수 모두를 대상으로 한 Simultaneous test가 선행되어야 한다. Simultaneous test 결과 해당되는 범주형 자료와 종속변수 간의 유의성이 확인되면 각각의 지시변수의 회귀계수를 이용하여 각 독립변수에 대응되는 범주의 평균값을 비교할 수 있다 (Montgomery, 2005).

ANCOVA 모형은 독립변수에 수치형과 범주형 자료가 함께 나타나는 경우 사용되는 모형이다. 이는 선형회귀 모형과 ANOVA 모형을 혼합한 것으로, 수치형과 범주형 자료 각각의

변수에 해당하는 회귀계수는 의미는 위에서 설명한 것과 동일하다.

앞에서 언급한 세 가지의 선형모형은 $\mathbf{X}$의 성격에 따른 구분이라 할 수 있다. 하지만 이들 세 모형은 진단하는 방법에 있어서 공통점을 보인다. 모형을 진단하는 방법은 잔차분석, leverage, Cook의 거리 등 여러 가지가 있다. Cook의 거리 (Cook, 1977)는 전체 자료에서 개개의 관측값을 제거하였을 때 회귀계수에 미치는 영향을 측정한 값이다. Cook의 거리가 작을 수록 모형을 구축하는데 영향을 덜 미치는 관측값이다. Leverage는 관측값에 대응하는 각 독립변수의 평균값을 살피면서 구축된 모형에 영향을 끼치는 자료를 판단하는데 사용된다. 만약 독립변수들의 평균값과 비교하여 차이가 많은 자료가 있다면 이는 이상치임을 판단하는 증거가 된다. 또한 이들 세 모형이 공유하는 성질(이 있는데 그것)은 오차항에 대한 분포적 가정으로 정의된다. 즉, 오차항들은 서로 독립이고 동일한 정규분포를 따른다. 흔히 오차항의 정규성, 독립성, 등분산성이라 부르며 모형을 추정한 후에 반드시 진단하는데 이를 잔차분석이라 한다 (Kutner *et al.,* 2005). 잔차분석 결과 오차항의 가정들이 만족되지 않는다면 모형 (1)의 회귀계수 추정치와 검정결과를 신뢰할 수 없다. 이러한 문제는 이상치가 존재하거나 또는 $\mathbf{y}$가 정규분포를 따르지 않기 때문에 발생한다. 전자의 경우라면 Cook의 거리나 leverage를 이용하여 이상치에 대한 조사를 통해 적절한 조치를 취한 후 다시 모형을 추정하는 것을 생각해 볼 수 있다. 후자의 경우에는 일반적으로 종속변수의 변환 예를 들면 Box-Cox변환 등을 고려해 볼 수 있다 (Box, George and Cox, 1964).

오차항의 독립성이 많이 훼손된 자료에 대해서는 위와 같은 조치로는 적절한 분석을 실시할 수 없다. 대표적인 예로 longitudinal data가 있는데 하나의 개체로부터 반복적으로 자료를 측정한 경우를 일컫는다 (Diggle *et al.,* 2000). 서로 다른 개체로부터 수집된 자료의 독립성은 일반적으로 받아들여진다. 하지만 동일 개체로부터 수집된 자료들의 독립성은 받아들여지기 어렵다. 이러한 자료들 간의 상관관계를 오차항의 공분산에 반영하여 분석하는 선형모형을 생각할 수 있는데 이를 혼합 모형이라 한다 (Pinheiro and Bate, 2004).

본 논문에서는 longitudinal data에 사용될 수 있는 혼합모형을 2장에서 간단히 소개하고 3장에서 정규성이 만족되지 않는 실제 자료에 대해 종속변수의 변환을 고려하여 분석한 결과 소개하도록 한다.

2. 혼 합 모 형

n개의 개체 중 i번째 개체로부터 관측된 자료 $\mathbf{y}_i = [y_{i1}, y_{i2}, \ldots, y_{im}]'$에 대해

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{a}_i + \boldsymbol{\epsilon}_i \tag{2}$$

와 같은 혼합모형을 생각한다. 여기서, i번째 개체로부터 관측된 자료는 m개이고, $\mathbf{a}_i$는 $q\times 1$ 랜덤 효과 벡터, $\mathbf{Z}_i$는 랜덤 효과에 대응되는 $m\times q$ 행렬, $\boldsymbol{\epsilon}_i = [\epsilon_{i1}, \epsilon_{i2}, \ldots, \epsilon_{im}]'$는 오차항을 나타낸다. $\mathbf{a}_i$와 $\boldsymbol{\epsilon}_i$은 서로 독립이고 그 (기대값, 공분산)이 각각 $(\mathbf{0}, \boldsymbol{\Sigma}_a)$와 $(\mathbf{0}, \boldsymbol{\Sigma}_\epsilon)$인 다변량 정규분포를 따른다고 가정하고, $\mathbf{a}_i \sim \mathrm{MN}(\mathbf{0}, \boldsymbol{\Sigma}_a)$와 $\boldsymbol{\epsilon}_i \sim \mathrm{MN}(\mathbf{0}, \boldsymbol{\Sigma}_\epsilon)$로 표기한다. 따라서 $\mathbf{y}_i$의 분포는 기대값은 $\mathbf{X}_i\boldsymbol{\beta}$고 공분산은 $\mathbf{V}_i = \mathbf{Z}_i\boldsymbol{\Sigma}_a\mathbf{Z}_i' + \boldsymbol{\Sigma}_\epsilon$인 다변량 정규분포다.

n개의 개체로부터 지속적으로 관측된 longitudinal data는 식 (2)와 같은 선형 혼합 모형을 이용하여 분석할 수 있다. 일반적으로 서로 다른 개체로부터 관측된 자료는 독립이라고 가정을 한다. 이 경우 전체 자료 $\mathbf{Y} = [\mathbf{y}_1', \mathbf{y}_2', \ldots, \mathbf{y}_n']'$의 모형은 $\mathbf{X} = [\mathbf{X}_1', \mathbf{X}_2', \ldots, \mathbf{X}_n']'$, $\mathbf{Z} = diag(\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_n)$, $\mathbf{a} = [\mathbf{a}_1', \mathbf{a}_2', \ldots, \mathbf{a}_n']'$, $\boldsymbol{\epsilon} = [\boldsymbol{\epsilon}_1', \boldsymbol{\epsilon}_2', \ldots, \boldsymbol{\epsilon}_n']'$라 할 때

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{a} + \boldsymbol{\epsilon} \tag{3}$$

으로 표현할 수 있으며, 그 분포는 기대값과 공분산이 $\mathbf{X}\boldsymbol{\beta}$와 $\mathbf{V} = diag(\mathbf{V}_1, \mathbf{V}_2, \ldots, \mathbf{V}_n)$인 다변량 정규분포다 (Pinheiro and Bate, 2004).

선형 혼합 모형이 식 (3)과 같이 제시되었다면 모수를 추정함으로써 longitudinal data를 분석할 수 있다. 공분산 $\mathbf{V}$가 주어질 때 $\mathbf{Y}$의 기대값을 구성하고 있는 $\boldsymbol{\beta}$의 GLS (generalized least square) 추정량은

$$\hat{\boldsymbol{\beta}}_{GLS} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y} \tag{4}$$

와 같다. 하지만, 많은 경우 $\mathbf{V}$도 추정해야 한다. 이 경우 식 (4)의 $\mathbf{V}$는 추정량 $\hat{\mathbf{V}} = \mathbf{Z}(\mathbf{I}_n \otimes \hat{\boldsymbol{\Sigma}}_a)\mathbf{Z}' + (\mathbf{I}_n \otimes \hat{\boldsymbol{\Sigma}}_\epsilon)$으로 대체된다. 일반적으로 $\boldsymbol{\Sigma}_a$와 $\boldsymbol{\Sigma}_\epsilon$을 추정하기위해 maximum likelihood (ML)과 restricted maximum likelihood (REML)를 이용한다 (Pinheiro and Bate, 2004).

추정된 모수를 바탕으로 통계적인 판단을 하기에 앞서 모형의 타당성을 확인하는 절차가 필요하다. 즉 잔차분석을 통해 $\boldsymbol{\epsilon}_i$들이 서로 독립이고 $\boldsymbol{\epsilon}_i \sim \mathrm{MN}(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$의 가정이 적절한 지를 확인해야 한다. 잔차 $\mathbf{e}_i$는 다양하게 정의되어 사용되고 있다. 보편적으로 알려져 있는 잔차는 $\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}$으로 각각의 관측된 자료값에서 자료 전체를 대표하는 평균값과의 차이로 정의된다. 하지만 이러한 정의는 식 (2)를 통해 알 수 있듯이 $\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}} = \mathbf{Z}_i\hat{\mathbf{a}}_i + \mathbf{e}_i$로 정확하게 $\boldsymbol{\epsilon}_i$

의 분포를 알려주지는 못한다. 이와 달리 longitudinal data의 각 개체는 저마다 구분되는 선형모형을 갖는다는 점을 반영하여 잔차를 $\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}} - \mathbf{Z}_i\hat{\mathbf{a}}_i$으로 정의할 수 있다. 본 논문에서는 오차 $\boldsymbol{\epsilon}_i$의 근사값으로써의 잔차를 $\mathbf{e}_i = \mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}} - \mathbf{Z}_i\hat{\mathbf{a}}_i$으로 정의한다. 이는 개체마다 관측된 자료값과 구축된 선형모형에 의해 적합된 값의 차이를 의미한다. 즉, 관측값에서 BLUE(best linear unbiased estimate)뿐만이 아니라 BLUP(best linear unbiased prediction)을 뺀 값이다. BLUP은 특히 선형 혼합 모형에서 랜덤 효과 부분을 예측하는데 사용되는 것으로 이는 모수 효과 부분의 BLUE와 동일한 의미이다 (Robinson, 1991). 이와 같이 정의된 잔차를 통해 모형의 타당성을 검정하는데 있어 ϵ_{ij}의 독립성, 등분산성, 정규성을 확인해야 한다.

식 (1)과 같은 랜덤효과가 없는 선형모형의 잔차 분석 결과 오차항의 가정이 성립되지 않는다면 생각해볼 수 있는 것으로 종속변수의 변환이 있다.

[표 1] TDO 임상실험 결과 표.

id	health	sex	visit	length	group
1	No	Male	0	7.3	1
1	No	Male	1	6.2	1
1	No	Male	2	8	1
1	No	Male	3	7.9	1
1	No	Male	6	7.6	1
1	No	Male	12	8.1	1
2	No	Female	0	5.3	1
2	No	Female	1	7.5	1
2	No	Female	2	6.9	1
2	No	Female	3	7.3	1
2	No	Female	6	7.2	1
2	No	Female	12	9.3	1
3	No	Male	0	5.8	1
3	No	Male	1	6	1
3	No	Male	2	6.5	1
3	No	Male	3	7	1
3	No	Male	6	9.1	1
3	No	Male	12	7.9	1
4	Yes	Female	0	3.3	1
4	Yes	Female	1	4.9	1
⋮	⋮	⋮	⋮	⋮	⋮

id	health	sex	visit	length	group
⋮	⋮	⋮	⋮	⋮	⋮
197	No	Female	6	9.8	2
197	No	Female	12	15.2	2
198	No	Male	0	4.2	2
198	No	Male	1	4.3	2
198	No	Male	2	7	2
198	No	Male	3	8.2	2
198	No	Male	6	8.1	2
198	No	Male	12	8.4	2
199	Yes	Male	0	6	2
199	Yes	Male	1	6.2	2
199	Yes	Male	2	4.4	2
199	Yes	Male	3	6.2	2
199	Yes	Male	6	8.9	2
199	Yes	Male	12	10.2	2
200	Yes	Male	0	6.4	2
200	Yes	Male	1	6	2
200	Yes	Male	2	6.6	2
200	Yes	Male	3	8.6	2
200	Yes	Male	6	6.8	2
200	Yes	Male	12	6.3	2

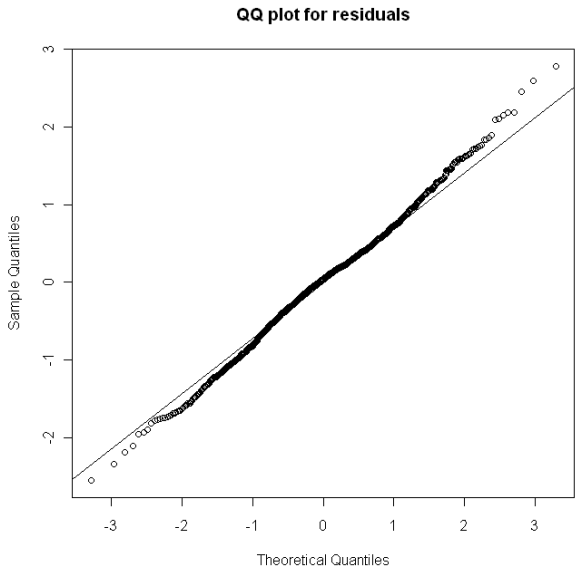
범용적으로 사용될 수 있는 종속변수의 변환인 Box-Cox변환은 $\mathbf{y}>0$에 대해

$$\begin{cases} \dfrac{\mathbf{y}^{\lambda}-1}{\lambda}, & \lambda \neq 0 \\ \ln \mathbf{y} \quad , & \lambda = 0 \end{cases} \tag{5}$$

으로 정의된다 (ref). 이러한 Box-Cox 변환을 식 (2) 모형의 $\mathbf{y}_i$에 적용하여 잔차의 가정이 성립하는 모형을 찾는 것을 시도해 볼 수 있다. 하지만 Box-Cox 변환이 항상 잔차의 가정 문제를 해결해 주지는 못한다. 즉 Box-Cox변환을 하였음에도 불구하고 $\boldsymbol{\epsilon}_i$의 정규성이 만족하지 않을 수 있다.

3. 예 제

앞 장에서는 longitudinal data에 적용할 수 있는 선형 혼합 모형에 대해 살펴보았다. 본 장에서는 실제 자료에 대해 혼합모형을 적용한 분석을 생각해 보도록 한다. 분석에 사용할 자료는 TDO(toenail dermatophyte onychomycosis)를 앓고 있는 200명의 미국인 환자를 대상으로 하여 임상실험을 한 결과로써 [표 1]과 같다. 각 환자를 경구복용약에 따라 랜덤하게 두 그룹(group)으로 나누어 환자가 치료를 받으면서 자라나는 정상적인 발톱의 길이(length)를 0, 1, 2, 3, 6, 12개월(visit)에 반복 측정한 임상실험 결과다. 두 그룹에 할당된 손발톱진균증을 치료하는 약은 각각 terbinafine과 itraconazole으로 terbinafine을 받은 그룹은 하루에 250mg을 투여하고(group=1), itraconazole을 받은 그룹은 하루에 200mg을 투여한다(group=2). 환자가 health club을 다니는 여부(health), 성별(sex) 또한 [표 1]결과에 제시되어 있다. [표 1]에 제시된 자료를 통해 환자의 발톱의 길이를 예측하는데 필요한 요인을 알아보고 또한 그 요인이 어떻게 영향을 주는지 선형 혼합 모형을 적용하여 통계분석을 실시한다. visit이 1이상인 환자 발톱의 길이를 종속변수로한 혼합모형을 생각한다. 독립변수로는 0개월일 때의 발톱의 길이(length0), sex, health, group을 이용하였다. 하지만 sex와 group은 유의한 결과를 보이지 않아 제외하였다. 이 경우 잔차에 대한 QQ plot을 작성해 보면 [그림 1]과 같다.

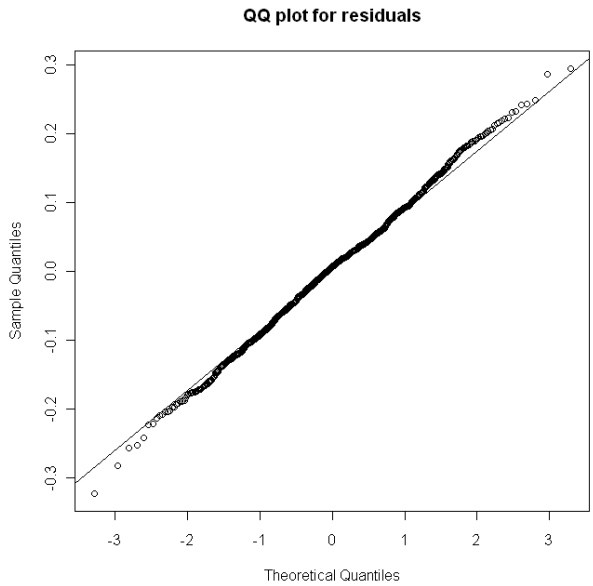


[그림 2] 종속변수 변환 전 잔차의 QQ plot.

QQ plot에 의하면 잔차의 분포는 정규분포에 비해 양쪽 꼬리부분이 두꺼운 것으로 이해될 수 있다. 정규성 검정을 위한 Shapiro-Wilcoxon test에서도 유의확률 P=0.02로 정규성이 만족되지 않은 일관된 결과를 확인했다. 뿐만 아니라 QQ plot 이러한 형태는 독립변수의 선택에 상관없이 그리고 이상치에 대한 조치를 취하여도 동일한 것으로 확인되었다. 따라서 종속변수의 변환을 고려했다. 하지만 longitudinal data에서 Box-Cox변환이 프로그램으로 아직은 구현이 되어 있지 않아 실사용에 어려움이 있다. 본 논문은 이러한 잔차의 분포적 불일치를 해소하기 위해 Box-Cox변환 대신 다음과 같은 변환을 제안한다.

$$\log\left(\frac{k+y_i}{k-y_i}\right)=\log\left(\frac{2k}{k-y_i}-1\right) \tag{6}$$

식 (6)을 사용하여 종속변수를 변환한 후에 다시 한번 혼합모형을 추정하였다. 여기서 k는 매우 작은 임의의 $\delta>0$에 대해 $\max(y_{ij})+\delta$로 정한다. 변수선택을 통해 여전히 동일한 독립변수들이 모형에 포함된 것을 확인하였다. 그러나 [그림 2]의 QQ plot을 살펴보면 잔차의 정규성이 만족됨을 알 수 있다. 이는 Shapiro-Wilcoxon test 에서도 동일하게 확인되었다 (P=0.65).



[그림 3] 종속변수 변환 후 잔차의 QQ plot.

참고문헌

[1] Box, George E.P. and Cox, D.R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society*, Series B 26, 211-246

[2] Cook, R.D. (1977). Detection of influential observations in linear regression. *Technometrics*, 19, 15-18.

[3] Diggle, P.J., Liang, Kung-Yee and Zeger, S.L. (2000). *Analysis of longitudinal data.* Oxford Science Publications.

[4] Kutner, M.H., Nachtsheim, C.J., Neter, J., and Li, W. (2005). *Applied linear statistical models* (5th edn). McGraw Hill.

[5] Montgomery, D.C. (2005). *Design and analysis of experiments* (6th edn.) John Wiley & Sons, Inc.

[6] Pinheiro, J.C. and Bates, D.M. (2004). *Mixed-effects models in S and S-Plus.* Springer.

[7] Robinson, G.K. (1991). That BLUP is a Good Thing: The Estimation of Random Effects. *Statistical Science,* 6 (1), 15-32.

[8] Verbeke, G. and Molenberghs, G. (2000). *Linear mixed models for longitudinal data.* Springer.

[9] http://www.stat.ncsu.edu/people/davidian/st732

[투고일: 2009. 2. 2] [심사일: 2009. 2.10] [게재확정일: 2009. 2.15]

Annals of Statistical Research
Vol. 5. No. 1, 2009, pp.55~74

The Current Status&Heed Analysis of After-school Activities in Elementary School[1)]

Soo-Dong Kim[2)]

Abstract

In order to implement after-school activities efficiently, from the perspective of principles&teachers in the elementary school, the current status and need analysis of after-school activities in elementary school is performed. Representative results are as followings. First, recognition improvement of after-school activities related personnels is needed. Second, excessive task burden of the teachers in charge of after-school activities is asked to be reduced. Third, expanding financial support is needed. Fourth, efficient implementation for programs is required. Fifth, rapid supply of competent lecturers should be implemented.

Key words : after-school activities, task burden on after-school activities, after-school activities program, completement lecturers for after-school activities.

1) This research is performed by revising & complementing 「Kim Soo-dong, Wang Seok-soon(2000). A Study on the Efficient Implementation of After-school Activities in Elementary & Secondary school. Korea Institute of Curriculum & Evaluation(KICE)」 'The current status and need analysis of after-school activities'

2) Associate Professor, College of Education, 707, Seokjang-dong, Gyeongju, Gyeongsangbuk-do, 780-714 Korea. E-mail: lyskhj1201@dongguk.ac.kr

Ⅰ. 조사방법

교과중심의 획일화된 학교교육을 극복하고, 학생들의 다양한 소질과 적성을 계발하기 위한 특기 · 적성교육을 효율적으로 실행하기 위하여 초등학교 특기·적성교육의 현황과 요구를 파악하고자 초등학교의 교장과 교사를 대상으로 설문조사를 실시하였다. 설문조사의 방법과 내용은 다음과 같다.

1. 대상

특기 · 적성교육에 대한 현황 및 요구조사 설문은 전국의 초등학교의 교장 및 교사를 대상으로 실시하였다. 표집방법에 있어서, 기존의 연구에서 표본수의 부족과 표집 지역이 편중되어 있어 전국적인 규모의 실태파악이 어려웠던 점을 극복하기 위하여, 본 연구에서는 전국의 초등학교에서 16개 시·도교육청별 학교 수 비율, 도시 및 농촌지역 등을 고려한 대규모의 비례유층표집을 사용하였다.

최종적인 표집은 230개교를 표집 하였으며, 표집된 학교의 교장 1인 및 교사 2인이 표집되었다. 따라서 대상자들의 배경변인이 비교적 고르게 분포되어, 내용적으로 세밀하고, 규모면에서 전국적인 파악이 가능하였다.

설문조사의 실시 기간은 2000년 4월 27일과 5월 13일에 걸쳐 이루어 졌고, 20일간 우편으로 설문지를 배포하여 회수하였다. 분석에 사용된 설문지의 배포, 회수, 회수율은 <표 Ⅰ-1>과 같다.

<표 Ⅰ-1> 초등학교 특기 · 적성교육에 관한 의견조사 설문지 배포, 회수, 회수율

대상	배포 (명)	회수 (명)	회수율 (%)
교장	230	120	52
교사	460	252	54
계	690	372	62
학교 수	230	130	56

2. 도구

본 조사에 사용된 도구는 '초등학교 특기·적성교육에 관한 의견조사'를 위한 설문지로서,

연구진이 제작하여 전문가의 협의를 거쳐 최종적으로 작성된 것이다. 설문지의 구성 내용은 총 5가지로 구성되어 있다. 첫째는 조사대상 인적사항, 둘째는 전반적인 운영현황, 셋째는 참여자들의 평가 및 기대, 넷째는 교장·교사가 본 운영상의 문제점 및 개선 방안, 다섯째는 운영비와 외부강사 관련 문제에 관한 의견이다. 구체적인 설문 내용과 해당 문항 번호는 <표 Ⅰ-2>와 같다(조사대상의 인적사항 분석은 생략함).

<표 Ⅰ-2> 특기 · 적성교육 의견 조사 설문지의 구성 내용

영역	내 용	문항번호	
		교장	교사
1) 조사대상 인적사항	가) 성별 및 연령	Ⅰ-1, Ⅰ-2	Ⅰ-1, Ⅰ-2
	나) 출신학교 및 학력	Ⅰ-3, Ⅰ-4	Ⅰ-3, Ⅰ-4
	다) 교직(장)경력	Ⅰ-5, Ⅰ-6	Ⅰ-5
2) 운영 현황	가) 학교의 특기 · 적성교육 실시 여부	7	
	나) 교사의 특기 · 적성교육 참여 현황		7
	다) 교사의 주당 특기·적성교육 담당시간		8
3) 참여자 평가 및 기대	가) 특기 · 적성교육 목적 인식		11
4) 교장과 교사가 본 문제점과 개선 방안	가) 특기 · 적성교육 운영상의 문제점	9-1~8	13-1~8
	나) 적절한 특기 · 적성교육 운영 장소	8	12
	다) 특기 · 적성교육 활성화를 위한 강좌개설 방안	10	14
	라) 특기 · 적성교육 활성화를 위한 수업운영 방안	11	15
	마) 특기 · 적성교육 활성화를 위한 시급한 개선영역	15-1~6	16-1~6
	바) 특기 · 적성교육 활성화를 위한 교육부 및 교육청의 시급한 지원 내용	18	19
5) 운영비 외부 강사 관련문제	가) 특기 · 적성교육 지도교사 적정 강사료	13	
	나) 특기 · 적성교육 수강료 및 강사료 결정권에 대한 인식	14	
	다) 외부강사 초빙에 대한 찬반	16	17
	라) 바람직한 외부강사 초빙 방식에 대한 인식	16-1~2	17-1~2
	마) 적절한 외부강사 요원에 대한 인식	16-3	17-3

Ⅱ. 조사결과 및 분석

1. 운영 현황

가. 학교의 특기 · 적성교육 실시 여부

학교장을 대상으로 현재 재직하고 있는 학교의 특기·적성교육 실시여부 및 실시시간에 대하여 조사하였다. 그 결과는 <표 Ⅱ-1>에서 보는 바와 같이, 초등학교의 경우 110명(95.7%)이 현재 자신이 재직하고 있는 학교에서 특기·적성교육을 실시한다고 응답하였다. 또한 현재 특기·적성교육을 실시하고 있는 학교들의 실시 시간에 대하여 살펴본 결과, 주당 3-4시간을 운영하는 학교들이 가장 높은 비율을 차지하고 있었으며, 그 다음으로는 1-2시간과 5-6시간 순으로 나타났다.

이와 같은 결과를 통하여 대부분의 초등학교들에서 특기·적성교육이 실시되고 있으며, 실시 시간은 주당 3-4시간이 가장 많음을 알 수 있다.

<표 Ⅱ-1> 초등학교의 특기 · 적성교육 실시 여부

구분	빈도(명) 및 비율(%)
실시하지 않음	5(4.3)
주당 1-2시간 운영	**20(17.4)**
주당 3-4시간 운영	**51(44.3)**
주당 5-6시간 운영	17(14.8)
주당 7-8시간 운영	5(4.3)
주당 9시간 이상 운영	17(14.8)
무응답	0(0)
총계	115(100)

나. 교사의 특기·적성교육 참여 현황

교사들을 대상으로 학교의 특기·적성교육 참여에 대하여 조사한 결과는 <표 Ⅱ-2>에서 보는 바와 같이, 담당하고 있지 않는 교사가 171명(67.9%)으로 나타났다. 그러므로 담당하고 있지 않는 교사가 상대적으로 많음을 알 수 있다.

이러한 결과는 앞에서 살펴본 실시학교와 참여 학생 현황과는 다소 차이가 있는 것으로 보인다. 즉 참여 학생이나 실시학교에 비해 담당교사의 비율이 상대적으로 낮음을 알 수 있는데, 이는 특기·적성교육을 현직 교사가 담당하기도 하지만 교외의 강사들을 활용하는 경우가 있으므로 이러한 결과가 나온 것으로 생각된다. 교사들의 참여 비율이 낮은 것은 교과보다는 예체능 분야의 프로그램이 많기 때문인 것으로 생각된다.

<표 Ⅱ-2> 초등학교 교사의 특기 · 적성교육 참여 현황

구 분	빈도(명) 및 비율(%)
담당하고 있다	73(29.0)
담당하고 있지 않다	**171(67.9)**
무응답	8(3.2)
총 계	252(100)

한편 '특기·적성교육을 담당하고 있지 않고 있다'라고 응답한 교사들을 대상으로 담당하지 않는 이유를 살펴본 결과, <표 Ⅱ-3>에서 보는 바와 같이 '담당교과와 관련된 특기·적성교육 활동이 없기 때문이다.'라고 응답한 비율이 가장 높은 것으로 나타났다. 다음으로는 '관련교과의 교사이나 담당하고 싶지 않다'고 응답한 비율이 높게 나타났다.

이러한 결과는 현직 교사들을 활용하기 위한 다양한 프로그램 및 참여기회의 부족과 특기·적성교육의 필요성과 교사의 적극적인 동참이 성공의 중요한 요인이라는 것에 대한 홍보와 교사들의 인식부족에 기인한 것으로 보인다. 또한 특기·적성교육에 대한 교사들의 자발적인 의욕부족 현상은 인센티브의 부족도 그 원인이라고 생각된다. 인센티브는 강사료 지급뿐 아니라 승진, 승급 등에 중요한 점수 부여 등 다양한 방법을 고려해야 할 것이다.

<표 Ⅱ-3> 초등학교 교사가 특기 · 적성교육을 담당하지 않는 이유

구 분	빈도(명) 및 비율(%)
담당교과와 관련된 특기 · 적성교육 활동이 없음	**88(51.5)**
담당교과가 관련 있으나 기회가 주어지지 않음	21(12.3)
담당교과와 관련이 있으나 교사 본인이 원하지 않음	50(29.2)
무응답	12(7.0)
계	171(100)

다. 교사의 주당 특기·적성교육 담당시간

교사들을 대상으로 현재 주당 특기·적성교육 담당시간에 대하여 조사하였다. 그 결과는 <표 Ⅱ-4>에서 보는 바와 같이, 39명(53.4%)이 주당 3-4시간을 담당하고 있다고 하였다.

이와 같은 결과를 통하여 대부분의 학교들에서 교사들은 주당 특기·적성교육을 3-4시간 담당하고 있음을 알 수 있다. 또한 5-6시간 담당하는 교사들도 15.1%가 있어서 정규수업 시간 외에 수업부담이 커서 정규 수업의 부실화가 우려되므로 담당 시간을 적정화해야 할 것이다.

<표 Ⅱ-4> 초등학교 교사의 주당 특기 · 적성교육 담당시간

구 분	빈도(명) 및 비율(%)
1-2시간 담당	17(23.3)
3-4시간 담당	**39(53.4)**
5-6시간 담당	11(15.1)
7시간 이상 담당	5(6.8)
무응답	1(1.4)
계	73(100)

2. 참여자 평가 및 기대

가. 특기·적성교육 목적 인식

특기·적성교육을 담당하고 있는 교사들을 대상으로 특기·적성교육을 통해 목적하는 바를 조사한 결과는 <표 Ⅱ-5>에서 보는 바와 같이, 예체능의 특기신장에 131명(52.0%), 특별활동의 연장에 67명(26.6%) 순으로 응답하였다.

이를 통하여 초등학교는 예체능의 특기신장을 주목적으로 하고 있음을 알 수 있다. 그러므로 상당히 유사성이 많은 특기·적성교육과 특별활동이 서로 연계되어 실시되는 것이 학생들의 특기·적성교육을 제대로 발굴하고 계발하는 방법임을 많은 교사들이 공감하고 있다는 것을 의미한다. 따라서 특기·적성교육과 특별활동의 연계방안을 적극적으로 모색해야 할 것이다.

<표 Ⅱ-5> 초등학교 교사의 특기 · 적성교육 목적 인식

구 분	빈도(명) 및 비율(%)
일반교과의 특기신장	13(5.2)
예체능의 특기신장	**131(52.0)**
특별활동의 연장	67(26.6)
정규교과 보충학습	5(2.0)
기타	24(9.5)
무응답	12(4.8)
총계	252(100)

3. 교장과 교사가 본 문제점과 개선 방안

가. 특기 · 적성교육 운영상의 문제점

초등학교 교장과 교사들을 대상으로 특기·적성교육 운영상의 문제점에 대하여 조사한 결과는 <표 Ⅱ-6>에서 보는 바와 같이, 교사의 업무 과중, 유능한 강사 채용의 어려움, 학급편성의 어려움 순으로 나타났다.

이와 같은 결과는 교사들과 학교장은 특기·적성교육에 관련된 업무가 늘어나는 것이 원활한 실시의 걸림돌임을 보여주는 것이다. 또한 유능한 강사 채용의 어려움이라고 답하여 학생과 학부모들의 다양하고 높은 요구를 수용할 수 있는 유능 강사의 채용이 어려움을 보여주는 것이다. 그리고 소수의 학생이 신청하는 경우 학생들의 부담이 늘어나 학급편성이 어렵다. 또 몇 달 동안 연속해서 수강해야 의미가 있는 강좌에서 일부 학생들이 중도에서 포기하면 강좌의 존속자체가 어렵게 된다.

<표 Ⅱ-6> 초등학교 교장과 교사가 인식한 특기 · 적성교육 운영상 문제점

구 분	유능한 강사 채용 어려움		학부모 참여 의지 부족		교사 업무 과중		담당교사와 비담당교사 간의 위화감		교재구성 어려움		학급 편성 어려움		학생 열의부족		지도교사 열의부족	
	빈도(명)	비율(%)	빈도(명)	비율(%)	빈도(명)	비율(%)	빈도(명)	비율(%)	빈도(명)	비율(%)	빈도(명)	비율(%)	빈도(명)	비율(%)	빈도(명)	비율(%)
매우 그렇다	51	13.9	31	8.4	111	30.2	24	6.5	29	7.9	47	12.8	25	6.8	4	1.1
그렇다	**162**	**44.1**	**133**	**36.2**	**158**	**43.1**	75	20.4	**136**	**37.1**	**137**	**37.3**	89	24.3	44	12.0
보통이다	90	24.5	102	27.8	66	18.0	**115**	**31.3**	123	33.5	96	26.2	**136**	**37.1**	**146**	**39.8**
대체로 그렇지 않다	53	14.4	82	22.3	25	6.8	104	28.3	62	16.9	61	16.6	89	24.3	130	35.4
전혀 그렇지 않다	10	2.7	18	4.9	5	1.4	40	10.9	10	2.7	17	4.6	21	5.7	39	10.6
무응답	1	0.3	1	0.3	2	0.5	9	2.5	7	1.9	9	2.5	7	1.9	4	1.1
계	367	100	367	100	367	100	367	100	367	100	367	100	367	100	367	100

따라서 특기·적성교육에 관련된 업무 부담을 줄여주고 유능한 강사를 채용할 수 있는 제도적인 장치와 노력들이 있어야 할 것이다.

나. 적절한 특기·적성교육 운영 장소

교장과 교사들을 대상으로 특기·적성교육 실시를 위한 적절한 장소에 대하여 조사한 결과는 <표 Ⅱ-7>에서 보는 바와 같이 교내나 교외를 정하여 두는 것보다 교내외에 걸쳐서 실시하는 것이 좋다는 의견을 보였다.

이러한 결과는 교장과 교사들은 학생들이 교내외에 걸쳐 특기·적성교육을 실시하는 것이 좋다고 생각하는 것이며, 또한 현재 교내에서 다양한 특기·적성교육을 실시하기 위한 시설이 부족함을 보여주는 결과이기도 하다. 교육부의 운영지침에서도 학교나 지역사회 시설 활용의 극대화를 권장하고 있고, 학교의 시설, 설비의 여건이 부실한 경우가 많기 때문에 이를 보완하기 위해서라도 외부의 시설을 적극 활용해야 할 것이다.

<표 Ⅱ-7> 초등학교 교장과 교사가 인식한 적절한 특기 · 적성교육 운영 장소

구 분	빈도(명) 및 비율(%)
교내	152(41.4)
교외	21(5.7)
교내외	180(49.0)
무응답	14(3.8)
총계	367(100)

다. 특기·적성교육 활성화를 위한 강좌개설 방안

학교장과 교사들을 대상으로 특기·적성교육을 활성화시키기 위한 강좌개설 방안에 대하여 조사하였다. 그 결과는 <표 Ⅱ-8>에서 보는 바와 같이, 162명(44.1%)이 학교여건에 맞게 특성화해야 한다고 하였다. 그 다음으로는 136명(37.1%)이 학생의 요구에 맞추어 개설해야

한다고 하였다.

이는 대체적으로 학생들의 요구를 반영하여 특기·적성교육의 프로그램이 개설되는 것이 특기·적성교육을 활성화시키는 길임을 교장과 교사들이 인식하고 있다는 것을 보여주는 것이다. 그리고 현실적 이유로 학생의 요구를 모두 수용하기 어려운 경우는 학교여건에 맞게 몇 가지 프로그램을 중심으로 특성화하는 방안도 적극 고려해야 할 것이다.

<표 Ⅱ-8> 초등학교 교장과 교사의 특기 · 적성교육 활성화를 위한 강좌개설 방안

구 분	빈도(명) 및 비율(%)
예체능 위주의 강좌개설	49(13.4)
일반교과 위주의 강좌개설	12(3.3)
학생의 요구에 맞춰 개설	136(37.1)
학교 여건에 맞게 특성화	**162(44.1)**
기타	7(1.9)
무응답	1(0.3)
총계	367(100)

라. 특기·적성교육 활성화를 위한 수업운영 방안

학교장과 교사들을 대상으로 특기·적성교육을 활성화시키기 위한 수업운영 방안에 대하여 조사하였다. 그 결과는 <표 Ⅱ-9>에서 보는 바와 같이, 192명(52.3%)이 가능한 다양한 수업방법이 필요하다고 생각하고 실행한다고 응답하였으며, 170명(46.3%)이 가능한 다양한 수업방법이 필요하다고 생각하지만 학교여건상 어렵다고 응답하였다.

그러므로 다양한 교수-학습방법 및 교수-학습자료에 대한 연구와 개발을 하여 이를 현장에 제공하는 방안도 고려해 볼 만하다. 이렇게 하여야만 특기·적성교육이 획일화된 교실 수업의 관행에서 실질적으로 벗어날 수 있을 것이다. 그리고 강좌당 인원수도 보통 20명 내외이기 때문에 다양한 교수-학습 방법을 사용하기에는 정규수업보다 훨씬 좋은 여건이다. 한편 정규 수업처럼 강의식으로 운영해야 한다고 응답한 교장이나 교사들은 1% 미만에 그치고 있다. 이것도 역시 특기·적성교육의 수업방식이 정규교과와는 달라야 한다는 생각을 교장과 교사들이 갖고 있으나 시설, 설비, 강좌당 인원수 등의 학교의 현실적 여건으로 어려움

을 겪고 있음을 보여주는 것이다.

<표 Ⅱ-9> 초등학교 교장과 교사의 특기 · 적성교육 활성화를 위한 수업운영 방안

구 분	빈도(명) 및 비율(%)
정규 수업처럼 강의 위주로 함	4(1.1)
다양한 수업방법이 필요하지만 학교 여건상 어려움	170(46.3)
가능한 다양한 수업방법이 필요하다고 생각하고 실행	**192(52.3)**
기타	1(0.3)
계	367(100)

마. 특기 · 적성교육 활성화를 위한 시급한 개선 영역

학교장과 교사들을 대상으로 특기·적성교육을 활성화시키기 위해 시급히 개선해야 할 영역에 대하여 조사한 결과는 <표 Ⅱ-10>에서 보는 바와 같이, 학급당 학생 수 축소, 외부강사 활용, 학생의 교과목 선택 보장, 능력별 반편성 순으로 높은 응답을 보였다.

이는 현재 우리나라 초등학교의 학급당 학생 수 과다가 특기·적성교육을 실시함에 있어 가장 큰 장애요인으로 작용하고 있음을 보여주는 결과이다. 또한 현직 교사보다는 학생들의 다양한 요구를 수용할 수 있고 싫증이나 지루함을 덜 느낄 수 있도록 적절한 외부강사의 활용이 시급히 요구된다는 것을 알 수 있다. 아울러 학생이 교과목 선택을 실질적으로 할 수 있도록 다양한 프로그램을 마련해야 할 것이다. 그리고 학생들의 개인차가 매우 크므로 능력, 적성, 흥미 등의 개인차를 고려하는 능력별 반편성이 필요함을 알 수 있다.

<표 Ⅱ-10> 초등학교 교장과 교사가 인식한 특기 · 적성교육 활성화를 위한 시급한 개선 영역

구 분	학생의 교과목 선택보장		학생의 지도 교사 선택 보장		외부강사 활용		능력별 반편성		학급당 학생 수 축소		강사료 인상	
	빈도 (명)	비율 (%)	빈도 (명)	비율 (%)	빈도 (명)	비율 (%)	빈도 (명)	비율 (%)	빈도 (명)	비율 (%)	빈도 (명)	비율 (%)
매우 찬성	**173**	**47.1**	72	19.6	**189**	**51.5**	**207**	**56.4**	**227**	**61.9**	104	28.3
약간 찬성	138	37.6	**137**	**37.3**	131	35.7	101	27.5	102	27.8	**150**	**40.9**
모르겠음	24	6.5	35	9.5	10	2.7	17	4.6	20	5.4	55	15.0
약간 반대	18	4.9	77	21.0	23	6.3	20	5.4	12	3.3	39	10.6
매우 반대	8	2.2	38	10.4	11	3.0	15	4.1	1	0.3	9	2.5
무응답	6	1.6	8	2.2	3	0.8	7	1.9	5	1.4	10	2.7
계	367	100	367	100	367	100	367	100	367	100	367	100

바. 특기·적성교육 활성화를 위한 교육부 및 교육청의 시급한 지원 내용

교장과 교사들을 대상으로 특기·적성교육의 활성화를 위하여 교육부 및 교육청이 시급하게 지원해야 할 내용에 대하여 조사한 결과는 <표 Ⅱ-11>에서 보는 바와 같이 재정적 지원을 가장 시급한 것으로 응답하였다.

이러한 결과는 우리나라 교육의 전반적인 문제를 반영하고 있는 것으로 교육재정의 확충을 통해 특기·적성교육에 필요한 재정을 충분히 지원하여 교육의 내실을 기할 수 있기를 바라는 것으로 볼 수 있다. 더불어 학생들의 다양한 요구를 반영하는 프로그램 개설 및 담당할 수 있는 유능한 강사의 수급이 필요함을 보여준다.

<표 Ⅱ-11> 초등학교 교장과 교사가 인식한 특기·적성교육 활성화를 위한 교육부 및 교육청의 시급한 지원 내용

구 분	빈도(명) 및 비율(%)
창의적 프로그램지원	113(17.3)
유능한 강사 발굴	141(21.6)
행정적 지원	46(7.0)
특기 적성교육 담당자	107(10.6)
재정적 지원	**234(35.8)**
기타 의견	12(1.8)
총계	653(100)

4. 운영비 및 외부강사 관련 문제

가. 특기·적성교육 시간당 적정 강사료

학교장을 대상으로 특기·적성교육의 지도교사에게 지급되는 시간당 적정 강사료에 대하여 조사하였다. 그 결과는 <표 Ⅱ-12>에서 보는 바와 같이, 28명(24.3%)이 2만원-2만5천원이나 1만원-1만5천원이 적정하다고 응답하였다.

<표 Ⅱ-12> 초등학교 교장이 인식한 특기 · 적성교육 지도교사 적정 강사료

구 분	빈도(명) 및 비율(%)
1만 원 이하	8(7.0)
1만원 - 1만 5천원 미만	28(24.3)
1만 5천원 - 2만원	21(18.3)
2만원 - 2만 5천원 미만	**28(24.3)**
2만 5천원 - 3만원 미만	18(15.7)
3만 원 이상	10(8.7)
무응답	2(1.7)
계	115(100)

나. 특기 · 적성교육 수강료 및 강사료 결정권에 대한 인식

교장을 대상으로 학교의 특기·적성교육의 수강료 및 강사료 결정권에 대한 인식을 조사한 결과는 <표 Ⅱ-13>에서 보는 바와 같다. 즉 50명(43.5%)이 학교운영위원회에서 결정하는 것이 바람직하다고 응답하였다. 그 다음으로 지목된 것은 학부모 대표이다. 학교운영위원회가 개별학교의 자율성을 높이고 학부모와 교장, 교사, 학생 및 지역사회 인사의 의견을 적절히 반영하기 위한 취지로 만들어진 기구임을 감안한 결과라고 볼 수 있다.

<표 Ⅱ-13> 교장의 특기 · 적성교육 활동 수강료 및 강사료 결정권에 대한 인식

구 분	빈도(명) 및 비율(%)
학교운영위원회	50(43.5)
교장단협의회	8(7.0)
교사협의회	12(10.4)
지역교육청	5(4.3)
교육활동 참여 학생들의 학부모 대표	37(32.2)
기타	3(2.6)
무응답	0(0)
총계	115(100)

다. 외부강사 초빙에 대한 찬반

교장과 교사를 대상으로 외부강사 초빙에 대한 찬반을 조사한 결과는 <표 Ⅱ-14>에서 보는 바와 같다. 667명(71.3%)이 찬성한다고 응답하였다.

이러한 결과는 교장과 교사들은 외부강사를 초빙하여 학교 특기·적성교육을 실시하는 것이 바람직하다고 생각함을 보여준다.

<표 Ⅱ-14> 교장과 교사의 외부강사 초빙에 대한 찬반

구 분	빈도(명) 및 비율(%)
찬성	667(71.3)
반대	100(10.7)
상황에 따라 다르다	137(14.7)
무응답	31(3.3)
총계	935(100)

라. 바람직한 외부강사 초빙 방식에 대한 인식

교장과 교사를 대상으로 바람직한 외부강사 초빙 방식에 대해 조사한 결과는 <표 Ⅱ-15>에서 보는 바와 같다. 즉 217명(69.6%)이 '강사은행 이용'을 찬성한다고 응답하였다. 다음으로는 '주변 사람들의 자문'이라고 응답하였다.

이러한 결과는 교장과 교사들은 강사은행을 이용하여 외부강사를 초빙하는 것이 바람직하다고 생각하고 있음을 보여주고 있다. 그 이유는 우선 편리하고, 어느 정도 공신력도 있다고 보기 때문인 것으로 생각된다. 그 다음으로 '주변 사람들의 자문'을 구하는 이유는 확실하고 믿을 수 있는 강사를 구할 수 있기 때문으로 생각된다.

<표 Ⅱ-15> 초등학교 교장과 교사의 바람직한 외부강사 초빙 방식에 대한 인식

구 분	빈도(명) 및 비율(%)
강사은행 이용	217(69.6)
주변 사람들의 자문	76(24.4)
기타	14(4.5)
무응답	5(1.6)
계	312(100)

마. 적절한 외부강사에 대한 인식

교장과 교사를 대상으로 적절한 외부강사가 누구라고 생각하는가에 대해 조사한 결과는 <표 Ⅱ-16>에서 보는 바와 같다. 즉 228명(34.2%)이 학원 강사라고 응답하였다.

이러한 결과는 교장과 교사들은 학원 강사가 가장 적절한 외부강사라고 생각하고 있음을 보여준다. 이는 특기·적성교육의 기능적 측면이 강조되고 있음을 보여주는 것이다. 뿐만 아니라 교장과 교사 역시 사교육을 담당하는 학원 강사들이 공교육의 부족한 점을 보완시켜 줄 수 있음을 인정하는 결과라고 볼 수 있을 것이다.

<표 Ⅱ-16> 교장과 교사의 적절한 외부강사에 대한 인식

구 분	빈도(명) 및 비율(%)
학부모 명예교사	106(15.9)
인근학교 교사	49(7.3)
학원 강사	**228(34.2)**
대학원생, 대학생	115(17.2)
기타	155(23.2)
무응답	14(2.1)
계	667(100)

Ⅲ. 실태 및 요구조사의 주요 결과 및 시사점

여기에서는 이상에서 제시한 실태 조사 결과 중 주요한 결과를 재논의하면서 보다 효율적인 특기 · 적성교육의 방안을 구안하기 위한 시사점을 도출하고자 한다.

첫째, 특기 · 적성교육 관련 당사자들에 대한 인식의 제고를 위한 다양한 접근이 필요하다.

특기·적성교육의 당사자는 정책의 입안자인 교육부, 교육청 관련 담당자 및 단위학교의 운영자로서의 교장과 교사, 그리고 교육의 수요자인 학부형과 학생 등이 있다. 이들 각 당사자들의 특기·적성교육에 대한 인식 제고는 교육의 효율적인 운영의 가장 중요한 열쇠가 될 수 있음을 실태 조사 결과를 통하여 유추할 수 있었다. 예를 들면 실태 조사 결과에서는 특기·적성교육에 대한 개선 방안으로 재정적 지원을 들고 있는데, 이러한 재정적 지원을 확충하거나, 재정적 지원을 시·도교육청별로 자율화, 개별화하기 위해서는 이에 대한 정책 당국자인 교육부와 시·도교육청 관계자들의 의지가 필요하다. 또, 조사결과에서 나타났듯이 특기·적성교육의 가장 중요한 문제점으로 지적된 '교사의 업무 과중'의 문제는 특기·적성교육 활동에 대한 중요성을 인식한 학교장의 행정처리 및 운영 편의를 위한 전폭적인 배려가 필요하며, 업무 과중 등의 문제점을 느끼고 있는 교사들에 대한 특기·적성교육 인식 제고를 위한 제반 조처가 뒤따라야 함을 시사 받을 수 있다.

둘째, 특기·적성교육 운영의 많은 책임을 지는 교사들의 업무 과중의 문제 해결을 위한 여러 가지 대책 마련이 필요하다.

특기·적성교육은 비정규 교육과정으로 운영되고 있으며, 실제로 이에 대한 교사의 책임을

어느 정도나 요구해야 하는가는 판단하기 어려운 상황이라고 할 수 있다. 실태조사 결과에서 볼 수 있듯이 교장과 교사들은 특기·적성교육의 가장 큰 난점을 '교사의 업무 과중' 문제로 지적하고 있다. 교사들은 방과 후에 다음 시간의 준비 등과 관련한 개인적인 시간을 가질 수 없다는 점에서, 또 외래 강사와는 다른 경제적 보상을 받거나, 대다수는 무보수로 특기·적성교육에 참여하고 있다는 점에서, 또 비정규 교육과정에서의 이중 근무에 따른 업무 과중, 각종 행정 처리와 관련한 업무처리상의 어려움 등에 대하여 부담감을 가지고 있음을 알 수 있다. 따라서 교직에 대한 봉사를 이유로 현재 별다른 대가없이 과중한 업무 부담을 주고 있는 현재의 운영상의 문제점을 개선하기 위한 다양한 방안이 필요함을 시사 받을 수 있다.

셋째, 특기·적성교육의 활성화를 위한 지속적인 재정적 지원과 재정지원에 따른 각종 업무 처리의 간소화를 통한 행정적 지원이 필요하다.

교장과 교사들은 특기·적성교육의 활성화를 위하여 교육부 및 교육청이 시급하게 지원해야 할 내용이 무엇인지에 대한 조사 결과에서 '재정적 지원'을 가장 시급한 내용으로 응답하였다. 특히 교육비의 학습자 부담 원칙이 어려운 지역의 학교일수록 재정적 지원에 대한 요구가 보다 높게 나타날 가능성이 높다.

한편, 특기·적성교육의 운영에는 복잡한 회계업무, 특히 국고 사용에 따른 행정처리 등의 여러 가지 회계 행정 업무, 그리고 외부 강사 관리, 프로그램 개설과 관리에 필요한 정기적인 관리 업무, 그리고 시설 운영에 따른 시설 관리 문제와 같은 각종 행정사항이 발생한다. 이러한 각종의 제반 업무는 특기·적성교육 담당자들에게 부담이 되고 있는 것이 사실이다. 따라서 이러한 업무 처리를 위한 행정상의 지원이나 관리비 부과 인정과 같은 각종의 지원책이 특기·적성교육에 대한 교사의 부담을 줄이며, 활성화에 기여할 수 있다고 사료된다. 따라서 현장 학교에서 보다 내실 있는 특기·적성교육이 운영되기 위해서는 지속적인 재정적 지원 및 그의 확충과 함께, 각종 행정 업무 처리와 관련한 행정적 지원이 병행하여 이루어져야 함을 시사해 주고 있다.

넷째, 특기·적성교육의 참여율을 향상시키며, 만족도를 향상시키기 위한 가장 중요한 요인은 프로그램 운영 측면에서의 개선과 관련된다.

초등학교 학교장과 교사들을 대상으로 특기·적성교육을 활성화시키기 위한 수업운영 방안에 대하여 조사한 결과를 보면, '가능한 다양한 수업방법이 필요하다고 생각하고 실행한다.'의 응답이 가장 많지만, 조사대상의 46.3%가 '다양한 수업방법이 필요하지만 학교 여건상

어려움을 겪고 있다'고 응답하는 것으로 나타나, 실제로 특기·적성교육 활동이지만 정규 수업과 다른 차별화된 교육을 하고 있지 못하다는 점을 알 수 있다. 특기·적성교육 수업 활동에서 나타나는 이러한 제한점은 특기·적성교육의 활성화에 걸림돌이 되고 있는 것으로 파악되고 있다.

다섯째, '원활한 강사 수급'과 관련한 강사문제의 해결은 특기·적성교육의 질 향상 및 프로그램 다원화의 원인이 되어, 특기·적성교육 활성화의 중요한 요인이 된다.

특기·적성교육 운영자인 교장과 교사들이 특기·적성교육 운영상의 문제점으로 '유능한 강사 채용'의 어려움을 주요 원인으로 지적하고 있다. 이러한 결과는 학생과 학부모들의 다양하고 높은 요구를 수용할 수 있는 유능 강사의 채용이, 다양한 특기·적성교육 프로그램의 개설에 영향을 주면서 동시에 질 높은 프로그램 운영에 관건이 되어 특기·적성교육 활성화에 크게 기여하는 것으로 설명할 수 있다.

따라서 앞으로 특기·적성교육을 수요자가 요구하는 '다양하되, 양질의 프로그램'으로 발전시키기 위해서는, '유능한 강사진 확보'가 관건이 되며, 이러한 강사진 확보를 위한 다양한 방안이 강구되어야 함을 시사 받을 수 있다.

여섯째, 특기·적성교육 활동 개선을 위한 여러 가지 정책 추진은 전국의 모든 학교에 적용되는 동일한 내용의 정책 추진보다는, 개별학교의 특성과 여건을 고려하여 다원화되고 융통성 있는 탄력적 정책 추진이 요구된다.

특기·적성교육 운영의 중심축인 개별 학교는, 개별 학교의 입지 지역, 학급수, 학생과 교사의 수, 공·사립의 구분, 학부형의 성향 등에 따라, 여건이 다양하다. 이와 같은 학교의 다양한 여건은 특기·적성교육 운영에 상당히 다른 양상을 가져오게 한다. 예를 들어 학교의 입지 지역이 도시이냐, 또는 농촌 지역이냐에 따라 수요자가 요구하는 프로그램이 다를 수 있고, 정부에서 지원해 주기를 원하는 지원책에 대한 요구가 서로 다를 수 있다. 도시 지역의 경우 비교적 강사 자원이 풍부하여 강사를 구하는 문제에서 덜 어려움을 겪는다면, 농촌 지역에서는 강사를 구하는 기초적인 문제에서도 어려움을 가지게 된다는 점이다. 비교적 중상류 지역의 학교에서 정부지원금을 받아 교육받아야 하는 학생이 별로 없기 때문에 정부 지원금이 크게 의미가 없는 반면에, 영세한 농어촌 지역에 입지한 학교는 정부의 지원금이 없는 경우 사실상 특기·적성교육을 수강할 수 있는 학생이 거의 없는 경우도 있다.

달리 말하면, 먼저 특기 적성 교육에 대한 정책 추진이나 이에 대한 적용은 중앙집권적인 정책을 통하여 획일적으로 적용되기 보다는 학교의 여건을 고려하여 다양한 집행을 할 수

있도록 개별적 학교 특성을 고려할 수 있는 시·도교육청, 또는 지역 교육청 수준으로 이관될 필요가 있다. 아울러 정책추진과 정책의 집행 요건이 개별 학교의 특성을 고려하여 보다 탄력적으로 적용될 수 있는 융통성이 필요하다는 점을 시사해주고 있다.

참고문헌

[1] 교육개혁위원회(1995). 『신교육체제 수립을 위한 교육개혁 방안』. 교육개혁위원회 제2차 대통령 보고서.

[2] 교육개혁위원회(1997). 『신교육체제 수립을 위한 교육개혁 방안(Ⅳ)』. 제5차 대통령 보고서.

[3] 교육부(1999a). 『1999 특기·적성교육 활동 운영계획』.

[4] 교육부(1999b). 『교육발전 5개년 계획(시안)』.

[5] 교육부(2000. 6). 『2000 특기·적성교육활동 운영 현황』.

[6] 김수동(1999). 개인차에 대응한 수업전략. 『수업연구』, 제27호. 전라남도교육과학연구원.

[7] 김수동, 왕석순(2000), 권양이. 『초·중등학교 특기·적성교육의 효율적 실행 방안 연구』. 연구보고 RRC 2000-4. 서울 : 한국교육과정평가원.

[8] 김재복(2000). 우리나라 초·중등학교 특기·적성교육의 발전 방향. 『특기·적성교육 활성화 방안 탐색 및 우수 사례 발표 세미나』. 연구자료 ORM 2000-12 세미나 자료집. 서울 : 한국교육과정평가원.

[9] 김재춘(2000). 초등학교에서의 효율적인 특기·적성교육 실행방안 토론. 『특기·적성교육 활성화 방안 탐색 및 우수 사례 발표 세미나』. 연구자료 ORM 2000-12 세미나 자료집. 서울 : 한국교육과정평가원.

[10] 부산 특기·적성교육 연구회(1999). 『부산시내 인문 고등학교 “특기·적성교육 실태” 분석 및 “개선 방안”』

[11] 왕석순(2000). 초등학교에서의 효율적인 특기·적성교육 실행 방안. 『특기·적성교육 활성화 방안 탐색 및 우수 사례 발표 세미나 자료집』. 연구자료 ORM 2000-12 세미나 자료집. 서울 : 한국교육과정평가원.

[투고일: 2009. 2. 11] [심사일: 2009. 2.15] [게재확정일: 2009. 2.15]

Annals of Statistical Research
Vol. 5. No. 1, 2009, pp.75~94

The effect of governance structure on performance in Franchise Organizations: Differences of Contractual and Relational Governance Structure

Sang-Kyu Lee[1], Young-Gook Kim[2]

Abstract

The purpose of this study is to examine the effect of governance structure on performance between franchisor and franchisee. The analysis of 202 franchisees indicated that relational governance as opposed to contractual governance is more effective in strengthening performance. This positive influence of relational governance is enhanced under low level of competitions.

This results suggest that the proficiency of relational governance may override the contractual mechanism in improving exchange and firm performance in franchise organization.

Keywords : Franchise organization, Contractual governance structure, Relationship governance structure, Exchange performance, Firm performance.

1) Doctorate Candidate, Dept. of Hotel and Tourism Mgt, Graduate School, Dongguk Univ.,Seoul. 100-715, email : rhgidf@hanmail.net
2) (Correspondance to) Professor, Dept. of Hotel and Tourism Mgt, Dongguk Univ. Gyongju, Korea 780-714. email : ygkim@dongduk.ac.kr

I. 서 론

프랜차이즈는 사업 확장과 창업활동에 매우 매력적인 조직형태이다. 이 조직형태는 기능적 측면에서 시스템적 표준화를 통하여 규모의 경제를 실현할 수 있고, 기업운영과 관련하여 소규모 창업자들에게 지역시장에 적응할 수 있도록 전문성과 적응력을 이용하는데 많은 장점을 가지고 있다(Kaufmann and Eroglu, 1999). 이러한 구조적 특성 때문에 많은 연구자들이 가맹본부와 가맹점 간의 관계에 대해 많은 관심을 가지고 있다.

경영자들은 프랜차이즈 본부와 가맹점과의 관계에 대해 이렇게 설명하고 있다. “프랜차이즈 본부와 가맹점 간의 관계는 결혼과 같다. 만약 누군가 이 관계를 맺기로 동의했다면 거기에는 많은 기본적 규칙이 있지만, 시간이 흐를수록 거기에는 해결해야 할 수 많은 일들이 존재한다(Bradach, 1997).” 이 문장은 두 파트너 사이의 가깝고도 먼 복잡한 관계를 적절하게 잘 표현하고 있다.

프랜차이즈 본부와 가맹점과의 관계에 대한 선행연구를 보면 크게 두 가지 주요한 흐름이 존재한다. 하나는 계약적 관계라고 보는 관점(Rubin 1978, Klein 1995, Brickley and Dark 1987, Agrawal and Lal 1995, Pazanti and Lerner, 2003)과 다른 하나는 관계적 관계라고 보는 관점(Bradach and Eccles, 1987 : Cochet, Dolmann, and Ehrmann, 2008 : 조규호, 전달영, 2003)이 존재한다. 계약적 관계란 본부와 가맹점 간의 관계를 계약적 개체로 파악하고 체인 본부가 가맹점을 관리하기 위해 이용하는 가장 중요한 수단이 계약파기이며 공식적이고 법적 구속력을 가지는 계약을 강조한다. 반면 관계적 관계는 본부와 가맹점 간의 관계를 신뢰와 협력을 바탕으로 파트너 간 강력한 상호작용과 지적교류를 활성화시켜주는 관계를 말한다.

따라서 두 연구결과에 근거한다면 두 관점이 모두 결과상으로 장단점을 내포하고 있으므로 두 관점 중 어떤 메카니즘이 프랜차이즈 관계에서 가맹점의 성과를 향상시키는데 효율적인지 파악할 수 없다.

따라서 본 연구는 앞서 제시된 두 가지 관점의 지배구조의 메카니즘인 계약적 관점과 관계적 관점에서의 효율성을 비교분석함으로서 어떤 형태의 지배구조가 본부와 가맹점간의 관계형성에 보다 더 효과적인지를 규명하는데 그 목적이 있다. 구제적으로 본 연구와 관련하여서 첫째, 일반 기업을 대상으로 한 선행연구결과에 의하면 계약적 지배구조가 기회주의 문제를 해결하고 관계를 안정화시키는데 기여하고 있다는 연구결과(Klein and Murphy, 1998)와 상호신뢰와 몰입이 기업의 성과에 보다 긍정적으로 영향을 나타내고 있다는 연구결과가 존재한다(Kaufmann and Stern 1998, Zaheer and Venkatraman, 1995). 따라서 두 유형의 지배

구조 중 어떤 메카니즘이 프랜차이즈 관계에서 가맹점의 성과를 향상시키는데 효과적인지를 규명하고자 한다.

둘째, 본 연구에서는 가맹점의 성과향상을 위해 두 지배구조를 병행해서 이용했을 때 프랜차이즈 관계에서 두 지배구조의 활용성이 상호 보완관계가 있는지를 규명하고자 한다.

마지막으로 본 연구에서는 동일한 프랜차이즈 가맹점간의 경쟁정도가 지배구조와 성과와의 관계에 영향을 미칠 것이라는 판단 하에 경쟁관계의 조절효과를 분석하고자 한다.

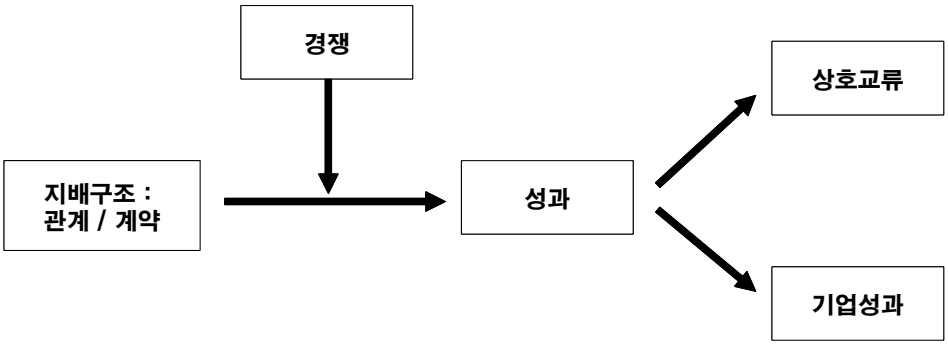
II. 개념적 모형과 가설설정

1. 개념적 모형

거래비용경제학(transaction cost economics)은 지배구조의 개념과 중요성에 대한 기본적 개념을 제공하고 있다. 이 연구의 일반적 명제는 경영자들은 알려진 상호위험(exchange hazards), 특히 전문적 자산투자나, 어려운 성과측정, 또는 불확실성에 관련된 상호위험요소에 대처하기위해 조직간 지배구조를 조정해나간다는 것이다(Williamson, 1981). 일반적으로 경영자들은 상호위험에 대응하여 복잡한 계약을 사용한다. 이 방법은 미래 상황에 대해 예측가능하게 하며 지속적인 해결방안을 제시해 줄 수 있고, 비예측적인 결과에 대처하기위한 구체적 절차를 제공해 줄 수 있는 반면에 이러한 계약은 사용하고 집행하는데 많은 비용이 들기 때문에 경영자들은 수직적 통합(vertical integration)을 선택한다(Macneil 1978).

그러나 많은 학자들은 거래비용경제학이 흔히 위험상황으로 지적된 상호교류에서 이러한 수직적 통합이나 명문화된 계약적 안전장치(contractual safeguard)의 효과에 대하여 과장하고 있다고 지적하고 있다(Zaheer, McEvily and Perrone, 1998). 이러한 관점에서 보면 신뢰와 같은 관계적 규범은 복잡하고 명문화 된 계약 또는 수직적 통합을 대신할 수 있는 대체수단으로 보고 있다(Bradach and Eccles 1989). 이런 근거로 신뢰와 규범적 행동은 계약이나 수직적 통합의 대안으로서 더 효과적이고 덜 비용이 드는 자율적인 안전장치로서 역할을 하고 있다고 판단된다. 따라서 본 연구는 앞서 제시된 두 가지 관점의 지배구조의 메카니즘인 계약적 관점과 관계적 관점에서의 효율성을 비교분석함으로서 어떤 형태의 지배구조가 본부와 가맹점간의 관계형성에 보다 더 효과적인지를 규명하는데 그 목적이 있다. 두 연구결과에 근거한다면 두 유형의 지배구조는 모두 결과상으로 장단점을 내포하고 있으므로 두 지배구조 중 어떤 메카니즘이 프랜차이즈 관계에서 가맹점의 성과를 향상시키는데 효과적인지 파악할 수가 없다.

따라서 본 연구에서는 본부와 가맹점간의 관계를 다음과 같은 개념적 모형으로 제안하여 가맹본부와 가맹점간의 효율적인 관계유지를 위한 지배구조를 규명하려고 한다. 구체적으로 1) 프랜차이즈 조직에서 가맹점의 성과에 대한 지배구조의 직접적 관계, 2) 관계적 지배구조와 계약적 지배구조 사이의 상호작용 관계 3) 지배구조와 성과간의 관계에 대한 가맹점 간 경쟁의 영향 등 이다.



<그림 1> 프랜차이즈 조직의 지배구조와 성과와의 관계

먼저, 기업 지배구조에 대한 선행연구의 결과에 따라(Williams 1985, Das and Teng 1998, Poppo and Zenger, 2002) 지배구조와 성과와의 관계를 측정하는데 두 차원의 지배구조를 사용하여 가맹점의 성과를 측정했다. 즉 계약적 지배구조는 공식적이고 법적 구속력 있는 협정이나 기업 내 파트너십을 지배할 수 있는 계약의 사용을 강조하며, 관계적 지배구조는 지배과정에 있어 상호 신뢰와 협력의 역할을 강조한다.

2. 관계적 지배구조와 성과와의 관계

거래비용경제학자들은 조직간 상호교류 지배구조의 효과에 대하여 지적하고 있다(Macneil 1978, Williamson, 1985). 이들에 따르면 조직간 상호교류는 사회적 관계에서 근거를 두고 있으며, 지배구조는 사회적 관계에서 나타난 가치나 합의 과정으로부터 나타난다고 하였다.

예를 들면 Poppo and Zenger(2002)는 관계를 가지고 지배하는 상호교류에서 책무나 약속, 기대의 이행은 규범에 있어서의 유연성, 결속, 정보교류를 증진시키는 사회적 과정을 통해서 일어난다고 지적하였다. 여기서 유연성은 비예측성사건에 적응하도록 도와주고, 결속은 상호조정을 통하여 공동의 행동에 대해 몰입하도록 하며 문제해결에 대한 쌍방의 접근방법을 모색하도록 한다. 정보공유는 파트너들이 서로 사적인 정보를 기꺼이 공유하도록 하기 때문

에 문제해결과 적응하는데 용이하게 한다. 파트너들이 그러한 규범에 몰입함에 따라 상호주의나 협력과 같은 행동으로 구체화되어진다.

관계적 지배구조는 상호신뢰와 협력을 강조한다(Roath Miller and Cavusgil 2002, Zaheer et al. 1998, Lee and Cavusgil, 2006). 이 연구들에 따르면 상호신뢰는 다음의 3가지 이유로 학습과 지적 교류를 용이하게 한다고 지적하고 있다.

첫째, 상호신뢰는 관련된 개인 간 강력한 상호작용을 용이하게 한다. 상호신뢰는 정보원, 정보의 가공방법 등을 제공 해준다.

둘째, 상호신뢰는 기회주의적 행위에 대한 두려움을 감소시켜줄 수 있다. 지식교류의 용이성과 유연성은 파트너 간 개방과 투명성의 정도에 달려있다. 의심은 지식을 공유하려는 의도를 감소시킨다.

셋째, 상호신뢰는 파트너들에게 정보와 학습노하우를 용이하게 해주는 지식공유기반을 구축하도록 도움을 준다(Zaheer and Venkatraman, 1995).

또한 상호신뢰가 지속적이고 효과적인 관계를 구축해준다는 연구도 있다. 예를 들면 Kaufman and Stern(1998)는 신뢰가 기업으로 하여금 관계적 지배를 통하여 자본구조의 종속성을 변화시키는데 유용하다는 것을 보여주었고 협상비용을 줄이고 전략성과를 높여준다고 주장하였다.

관계적 측면에서 보면 신뢰는 향후 상호교류에서 설득과 격려를 바탕으로 한 중요한 역할을 한다. 그래서 신뢰 기반형 파트너십은 안정적인 교류관계를 구축하고 나아가서는 상호교류 성과에 영향을 미치고(Bradach and Eccles 1989, Zaheer, et al., 1998). Ivens and Blois(2004), 협력적 행동을 유발시키며 불확실성을 감소시킨다고 하였다.

조규호, 전달영(2003)의 국내 프랜차이즈 연구에서도 가맹점과 체인본사 간의 거래관계에서 신뢰가 관계몰입에 유의하게 영향을 주는 것으로 분석되었다. 따라서 본부와 가맹점간의 관계에서 다음과 같이 제안한다.

가설1 : 프랜차이즈 조직의 관계적 지배구조는 가맹점의 상호교류 성과에 긍정적인 영향을 미칠 것이다.

가설2 : 프랜차이즈 조직의 관계적 지배구조는 가맹점의 기업성과에 긍정적인 영향을 미칠 것이다

3. 계약적 지배구조와 성과와의 관계

계약적 지배구조는 공식적이고, 법적 구속력을 가지는 합의나 기업 간 파트너십을 통제할 수 있는 계약의 이행을 강조한다. 공식계약은 미래 특별한 조치를 실행하기 위한 약속이나 의무를 나타냈다(Poppo and Zenger, 2002).

일반적으로 계약이 복잡하면 할수록 분쟁에 대한 약속, 책무, 절차에 대한 세부사항이 더 많아지고 복잡해진다. 예를 들면 복잡한 계약은 수행해야 되는 책임과 역할이 자세히 구체화되거나 감시절차나 위반사항을 구체적으로 제시하고, 더 중요한 것은 전달된 결과를 통제하도록 되어 있다. 거래비용 경제학의 논리에 따르면 경영자의 목표는 요구된 양, 가격, 그리고 공급자의 서비스 품질의 전달을 보장할 수 있도록 최소비용으로 지배구조를 개발하는 것이다. 따라서 경영자는 다양한 서비스에 부합하는 상호조건에 적합하도록 지배구조를 선택한다. 상호위험(exchange hazards)이 증가할수록 계약적 안전장치(contractual safeguard)가 필요하다. 즉 비용을 최소화 할 수 있고 그러한 위험이 발생할수록 성과는 감소한다(Williamson 1985, Klein, 1995).

거래비용경제학자들은 통상 계약적 안전장치나 수직적 통합이 필요한 상호위험(exchange hazards)의 3가지 요소를 지적하고 있다. 즉 자산의 특유성(asset specificity), 측정의 어려움(measurement difficulty), 그리고 불확실성(uncertainty)의 3가지이다(Heide 1994). 이 세 가지 요소가 더욱 복잡한 계약을 촉진시킬 수 있다는 것이다. 이와 관련하여 복잡한 계약은 추가적인 계약비용이 요구된다. 이 추가비용요소를 무시한다면 상호위험상황에 대응하는 효율적 수단으로서 복잡한 계약을 이용하는 것이 오히려 상호교류를 활성화시킴으로서 성과를 향상시킬 수 있다. 오히려 복잡한 계약이 두 파트너 간 기회주의 행동을 사전에 방지할 수 있다. 계약적 지배구조는 파트너십에 대한 위험요소를 줄이고 지식교류를 용이하게 하며, 상호교류나 기업성과를 향상시킨다. Williamson(1985)에 따르면 거래의 능률성과 기회주의(opportunism)에 대한 특유의 재산 보호의 균형을 맞추기 위해 기업은 위계적이고 시장지향적인 복합적인 지배구조를 이용할 것이다. 시장 지향적 거래는 자발적인 경제적 실체 간의 교환관계를 포함하고 있으며 효율적인 계약방법으로서 역할을 하고 있다. 고전적인 계약은 기업이 스스로 파트너의 기회주의로부터 보호하기위한 효율적인 안전장치 역할을 한다. 명문화된 계약은 거래관계가 구속력을 가질 수 있다는 것을 확인시켜준다. 그래서 기업은 공유된 정보의 양과 유형을 통제하고, 지식 교류 시 파트너가 의도한 범위를 초과할지 모른다는 위험요소를 줄이고 미래기업간 신뢰구축을 위한 기반을 마련하기 위해, 영향력 있는 계약협정을 선호한다.

계약적 지배구조는 갈등관리를 위한 수단을 제공할 수도 있다(Klein and Murphy, 1998). 만약 불화가 발생하면 계약조항은 무엇이 법을 위반했는지 판단할 수 있다. 또한 체인 본부가 가맹점을 관리하기 위해 이용하는 가장 효율적인 수단이 계약파기라고 주장하는 학자들도 있다(Rubin 1978, Klein 1995, Brickley and Dark, 1987).

따라서 다음과 같이 가설을 본부와 가맹점 간의 관계에서 설정 할 수 있다.

가설3 : 프랜차이즈 조직의 계약적 지배구조는 가맹점의 상호교류 성과에 긍정적인 영향을 미칠 것이다.

가설4 : 프랜차이즈 조직의 계약적 지배구조는 가맹점의 기업 성과에 긍정적인 영향을 미칠 것이다.

4. 상호작용효과와 성과와의 관계

앞에서 관계적 지배구조와 계약적 지배구조는 장단점을 가지고 있다. 따라서 프랜차이즈 조직에서 이 두 메커니즘을 결합한다면 서로의 장단점을 보완할 수가 있어서 본부와 가맹점 간의 효율적 관계유지에 긍정적으로 영향을 미칠 것으로 판단된다. 과연 관계적 지배구조와 계약적 지배구조의 결합이 단일지배구조를 이용하는 것보다 더 높은 성과에 영향을 미칠 것인가? 대답은 단순하지가 않다.

관계적 지배구조는 계약적 지배구조를 보완해줄 수도 있다.

첫째, 선행연구에 따르면(Jones, et al., 1997) 거래비용경제학은 위기상황에서 계약적 안전장치의 효과를 과장하고 있다고 지적하고 있다. 그러나 복잡한 계약은 고비용을 전제로 하고 있어서 신뢰와 규범을 강조하는 관계적 지배구조는 계약적 지배구조보다 덜 비용이 들고 효율적이고 자율적인 안전장치로서 역할을 할 수 있다(Roath, et al., 2003).

둘째, 계약서는 쌍방의 요구사항을 정하고 관계를 유지하기 위한 기본적 틀만을 제시한다. 모든 실제상황을 계약으로 구체화할 수 없다. 예기치 않은 불안요소가 나타나면 계약은 파트너 간 관계에서 영속성을 유지할 수 없다. 따라서 관계적 지배구조는 변화와 갈등이 발생 시 상호주의와 영속성을 제공함으로서 계약적용의 한계에 대한 필요한 완충장치의 역할을 할 수 있다.

계약적 지배구조는 관계적 지배구조를 보완해 줄 수 있고, 불확실성을 구체화시켜줌으로서 양자 간의 갈등요소를 해소시켜줌으로서 관계적 지배구조를 보완해 준다. 양자 간 잘 다듬어지고 협의된 장기계약은 궁극적으로 상호교류를 원활하게 함으로써 상호불만요소를 사전에

제거해 준다.

즉, 프랜차이즈 조직의 계약적 지배구조는 명문화된 자세한 계약관계, 처방 그리고 논쟁해결 방법을 제공하고, 관계적 지배구조는 신뢰, 관계적 규범의 유연성, 상호주의, 관계지속성을 향상시켜 줄 수 있다. 이 두 지배구조의 메커니즘의 결합은 가맹점의 상호작용을 원활하게 해주며 기업 성과를 향상시킬 것이다. 이 점을 토대로 다음과 같이 본부와 가맹점간의 관계에서 가설을 세울 수 있다.

가설5 : 프랜차이즈 조직의 관계적 지배구조와 계약적 지배구조의 결합은 가맹점의 상호작용 성과를 증가시킬 것이다.

가설6 : 프랜차이즈 조직의 관계적 지배구조와 계약적 지배구조의 결합은 가맹점의 시장성과를 증가시킬 것이다.

5. 가맹점 간 경쟁관계의 조절효과

자원기반이론(resource-based theory)에 따르면 프랜차이즈 초기에는 생산, 마케팅, 조직관리에서 규모의 경제를 실현할 수 있는 최소규모(critical mass)를 확보하고 있지 못하기 때문에 경제적 효율성이 떨어질 수밖에 없다. 따라서 프랜차이즈 시스템이 발전하기 위해서 본부는 최대한 빠른 시기에 가맹점 추가 모집을 통한 성장전략을 택하여 규모의 경제를 실현함으로서 체인의 경쟁력을 구축하려고 한다(Shane 1996). 이 때 본부와 가맹점 간 갈등이 불가피하다.

체인 내 가맹점 간 경쟁관계는 시스템의 계속된 성장과 더불어 확연하게 나타난다. 특히 성숙단계에 들어서게 되면 가맹본부는 가맹점의 추가 모집을 통하여 이미 개발된 경제적 잠재력을 지역적 접근성을 통하여 구축하려고 한다. 특히 본부는 가맹점간의 밀집화 전략(clustering)을 통하여 수평적 통합을 이룩하도록 한다(Stassen and Mittelstaedt, 1995).

동일지역 내 밀집화 전략은 가맹점 간 치열한 경쟁을 초래하게 되고, 결국 중심 가맹점의 시장규모가 감소하게 되고, 그 결과 투자 회수율에도 영향을 줄 수 있다. 가맹점의 시장이 줄어들고 가격인하 압력과 함께 체인 내 경쟁관계는 매출액 감소로 이어질 수 있다.

결국 이 치열한 경쟁은 시장규모를 감소시키고 가맹점은 제살 깍아먹기를 초래한다. 이는 가맹점의 불만으로 이어지고, 성과에도 부정적으로 영향을 미칠 수 있다.

따라서 다음과 같은 가설을 세울 수 있다.

가설7 : 가맹점간 경쟁수준은 관계적 지배구조와 상호교류 성과와의 관계를 조절할 것이다. 구체적으로 두 개의 긍정적 관계는 가맹점간 경쟁관계가 낮을 때 강화 될 것이다.

가설8 : 가맹점간 경쟁수준은 계약적 지배구조와 기업 성과와의 관계를 조절할 것이다. 구체적으로 두 개의 긍정적 관계는 가맹점간 경쟁관계가 낮을 때 강화될 것이다.

III. 연구방법

1. 표본

본 연구에서는 서울, 경기지역의 다양한 프랜차이즈 업체의 가맹점을 대상을 실시하였다. 가맹점의 지배구조와 성과와의 관계를 측정하기 위하여 업체선정은 기업규모와 설립년도를 반영하여 선정하였다. 이는 프랜차이즈 시스템을 잘 갖춘 업체가 본부와 가맹점간 분석 연구의 왜곡이 덜 일어날 수 있다는 선행연구결과를 반영하였다(오세조·김상덕·오일두, 2003).

한국프랜차이즈 협회 홈페이지의 분류기준에 따라 업체를 선택하여 홈페이지에서 주소, 가맹점의 수, 설립년도 등을 파악하여 가맹점을 대상으로 설문방식을 실시하였다. 총 208개의 가맹점의 점주 또는 점장을 대상으로 실시하였다. 총 202부를 최종적으로 분석에 이용하였다. 표본을 보면 외식업 69% 로 가장 많았으며, 이 미용업 등 서비스업 14.6% 편의점 등 소매업 13.5% , 학원 및 기타 17.5%로 구성되었다. 남성이 86.7로 여성보다 많았으며 기혼자가 73.2% 이며, 학력을 보면 대졸이 57.7 % 로 가장 많고, 고졸이 32%로 뒤를 이었다. 한편 연령은 40-50세가 33.2%, 한 가맹본부 당 평균 가맹점 수는 12.4개 분포되고 있으며, 평균가맹점 가입 년 수는 1년 5개월이다.

2. 측정도구의 신뢰성과 타당도

본 연구에서 이용된 변수의 조작적 정의와 측정도구의 출처는 다음과 같다.

지배구조는 선행연구에 따라(Poppo and Zenger, 2002, Roath, et. al., 2002) 지배구조를 두 개의 차원으로 분류하였다. 관계적 지배구조는 신뢰와 협력, 갈등해소를 강조한다. 계약적 지배구조는 공식적이고 법적 구속력을 가지는 합의나 파트너 간 관계를 통제할 수 있는 계약을 강조한다. 본 연구에서는 관계적 지배구조는 Cochet, et al.(2008)의 연구에서 사용된 측정도구를 이용하였으며 계약적 지배구조도 이전연구에서 사용되었던 측정도구를 이용하였다

(Macnei, 1978, Yikuan and Cavusgil, 2006). 관계적 지배구조는 본부와의 관계에서 신뢰와 협력, 그리고 갈등해소 등을 질문하는 6문항(각2문항)으로 구성되었으며 계약적 지배구조는 본부와의 관계에서 공식적 계약, 법적 책무, 그리고 계약의무 준수, 구속력 등을 질문하는 5문항으로 구성되었다.

성과는 두 차원으로 측정하였다. 지배구조 관련 선행연구에서 보면 성과 측정은 지배구조의 효율성에 초점을 맞추어져 있다(Uzzi 1997, Artz and brush 2000, Wu and Cavusgil, 2006). 따라서 성과 측정을 상호교류 성과(exchange performance)와 기업 성과(firm performance)의 두 개의 차원으로 분류하였다. 상호교류성과는 지배구조의 효율성과 관련하여 파트너 간 상호교류를 통한 만족으로 측정하였다. 이러한 상호교류의 성과는 양자 간의 파트너십의 가장 중요한 결과는 상호만족에 있기 때문이다. 따라서 상호교류성과는 상호교류의 강화와 안정성, 지적 교류에 대한 만족정도로 측정하는 문항으로 총 4문항이며, 기업 성과는 가맹점의 매출 성장, 상품개발, 수익성 등을 질문하는 문항으로 총 4문항으로 구성 측정하였다. 모든 측정은 리커트 5점 척도를 이용하였다.

경쟁관계는 동일체인 내 가맹점 간 경쟁의 강도를 특정하기 위하여 가맹점들에게 관내 가맹점수가 적정여부를 측정하였다. 이 문항은 지역 내 동일 브랜드의 경쟁수준에 대하여 가맹점이 느끼는 정도를 질문하는 형태로 Cochet, et. al.(2008)의 연구에서 이용되었던 1개의 문항을 사용하였다.

본 연구에서 사용된 통제변수는 프랜차이즈관련 선행연구에서 규모(가맹점수), 가맹사업기간(가입연도), 로얄티 비율 등이 거론되었다(Cochet, et al., 2008). 이 중에서 로얄티 비율은 김응수·임영균(2006)에 따르면 우리나라의 경우 상당수가 로얄티를 부과하지 않고 있고(65%), 부과하고 있는 경우도 정액제를 실시하는 가맹본부의 비율이 높아 로열티 비율의 정확한 산정이 어려워서 본 연구에서는 제외하였다. 규모는 본부의 홈페이지 등을 이용하여 파악하였고, 가맹사업기간은 시스템에 가입한 년도는 질문하는 문항으로 측정하였다. 본 연구에서 사용된 변수에 대한 신뢰도는 Crombach의 α값으로 검증하였다. 관계적 지배구조는 .788 계약적 지배구조는 .776, 상호작용 성과는 .811, 기업 성과는 .867 으로 나타났다.

변수들의 타당도를 검증하기 위하여 요인분석을 이용하였다. 요인추출방식으로는 주성분분석방법을, 요인회전방식으로는 직각회전방법을 이용하였다. 요인분석에 이용된 항목은 19개인데 분석에 사용된 표본은 202부이기에 기준(5배)을 충분히 충족시킨다(Hair et al. 1995). 분석에 이용된 변수들에 대한 요인분석 결과는 <표1>에 나타나고 있다.

<표 1> 측정변수의 요인분석결과

문 항	관계적 지배구조	계약적 지배구조	상호교류 성과	기업성과	Cronbaha α
rg 1	.764				.788
rg 2	.657				
rg 3	.793				
rg 4	.822				
rg 5	.803				
rg 6	.811				
cg 1		.790			.776
cg 2		.833			
cg 3		.792	.		
cg 4		.498			
cg 5		.627			
ap 1			.758		.811
ap 2			.799		
ap 3			.792		
ap 4			.537		
mp 1				.755	.867
mp 2				.783	
mp 3				.589	
mp 4				.705	
아이겐값	4.367	3.998	2.747	1.520	
설명비율(%)	32.655	12.442	10.445	9.863	
누적설명비율(%)	32.652	44.094	54.536	63.399	

요인분석결과 아이겐 값이 1이 넘는 요인의 수는 모두 4개로 나타났고 총 분산가운데 63.399%를 설명하였으며 이는 본 연구에서 측정하고자하는 항목들이 모두 유의한 것을 판단된다.

IV. 분석결과

1. 변수들에 대한 기초분석

본 연구에서 사용된 변수들의 평균, 표준편차 및 상관관계 분석결과는 다음과 같다. 표2에

서 보면 상관관계를 통해서 변수 간 대략적인 관계를 파악할 수 있다. 회귀분석에서 다중공선성의 문제는 상관관계계수가 .8 보다 커야 생각해볼 수 있는데(Hair, et al., 1998), 본 연구에서 얻어진 상관관계의 수치로서는 다중공선성 문제를 제기할 수 있는 수준에는 못 미치는 것으로 나타났다.

<표 2> 변수 간 평균, 표준편차, 상관관계

구분	평균	표준편차	1	2	3	4	5
관계지배구조	3.45	.5655					
계약지배구조	3.12	.6378	.338**				
상호교류성과	3.36	.5767	.406**	.123*			
기업성과	3.31	.5672	.221**	.172*	.299**		
경쟁수준	3.12	.5764	-.131*	.033	-.134*	-.074	

주) * p<.05, ** p<.01 모든 변수들은 5점 척도로 측정되었음

본 연구에서 설정한 가설검증을 위해 <표 3>과 같이 다중회귀분석을 실시하였다. <표 3>에서 주효과 모형을 보면 단지 관계적 지배구조만이 두 성과에 긍정적으로 영향을 미치고 있음을 알 수 있다. 다시 말하면 관계적 지배구조는 상호교류성과(β=.443, p<.001)와 기업성과(β=.232, p<.01)에 긍정적으로 영향을 미치는 것으로 나타나고 있으나, 계약적 지배구조는 두 성과에 관계가 없는 것으로 나타났다. 따라서 가설1,2 은 지지되고 있으나 가설3,4는 지지되지 못했다. 이 연구결과는 프랜차이즈조직에서 본부와 가맹점간의 효율적인 관계 유지를 위해서는 관계적 지배구조가 훨씬 효과적이다. 구체적으로 말하면 관계적 지배구조가 계약적 지배구조 보다 가맹점의 상호교류성과와 기업성과에 더 효과적이고 더 영향을 미치는 것으로 나타났다.

또, 관계적 지배구조와 계약적 지배구조 간 상호작용효과를 측정하였다. 측정결과 가설5,6과는 반대로 계약적 지배구조는 가맹점의 성과를 향상시키기는데 관계적 지배구조를 보완하고 있지 않았다. 즉, 상호교류성과에 대한 상호작용 효과는 상호교류성과(β=-.222, p<.01)와 기업성과(β=-.152, p<.01)와 유의적으로 부정적 관계를 가지고 있다. 이 연구결과는 계약적 지배구조가 관계적 지배구조를 보완하는 것이 아니라 부정적인 영향을 일으키고 있다고 판단된다. 이는 계약적 지배구조가 관계적 지배구조를 방해하고 있다고 추측된다. 따라서 이 가설은 지지되지 못했다.

<표 3> 지배구조와 성과에 대한 다중회귀분석결과

변 수	상호교류 성과		기업 성과	
	모형 1	모형 2	모형 3	모형 4
통제변수:				
가입년수	.100	.101	.113	.165*
가맹점수	.088	.123	.075	.151*
경쟁관계	-.183*	-.187*	-.136	-.188*
주효과:				
관계적 지배구조	.423***	.443***	.211**	.232**
계약적 지배구조	.077	.079	.093	.121
상호작용효과:				
관계적지배구조*계약적지배구조		-.162*		-.152*
R2	.239	.277	.211	.221
F	10.262***	14.324***	7.555***	9.799***

주) * p<.05, ** p<.01, *** p<.001

2. 가맹점의 경쟁에 따른 조절효과

<표 4>는 가맹점의 경쟁수준에 따른 조절효과를 분석하였다. 먼저, 가맹점의 경쟁수준을 중앙값을 이용하여 고/저로 나누어 지배구조와 성과와의 관계를 분석하였다. 분석결과 가맹점의 성과에 대한 지배구조의 영향은 경쟁정도의 고저에 따라 달라질 수 있다는 것을 보여주고 있다. 즉 상호교류성과에 대한 관계적 지배구조의 영향은 높은 수준의 경쟁수준(β=.243, p<.05) 보다 낮은 경쟁수준에서(β=.543, p<.001) 더 강력한 긍정적 영향력을 보여주고 있다(F=3.055, p<.05). 그러나 기업성과에 대한 관계적 지배구조의 영향은 유의적이지 못하다.

가맹점의 성과에 대한 계약적 지배구조의 영향은 경쟁수준에 따라 차이가 없는 것으로 나타났다. 따라서 가설7.8 은 부분 지지되었다. 연구결과는 가맹점간 경쟁수준이 높을수록 본부와의 신뢰관계에 부정적으로 영향을 미치게 되고 결국 성과(상호교류)에도 부정적으로 영향을 미치는 것으로 나타났다. 연구결과는 가맹점간 경쟁수준의 적정성이 신뢰관계에도 영향을 미치고 나아가서는 상호교류 성과에도 긍정적 관계가 있다는 것을 보여주고 있다.

<표 4> 가맹점의 경쟁정도에 따른 조절효과 분석

변 수	종속변수 : 성과			
	상호교류 성과		기업 성과	
	저	고	저	고
통제변수: 가맹점수 가맹년도	.098 .164*	.161* .111	.160* .101	.103 .111
독립변수: 관계적 지배구조 계약적 지배구조	.543*** .121	.243* .120	.326** .122	.311** .121
F	3.055*		1.554	

* p<.05, ** p<.01, *** p<.001 고/저는 가맹점간 경쟁수준이다..

V. 결론, 시사점 및 연구의 한계

본 연구결과는 프랜차이즈조직에서 본부와 가맹점 간의 관계를 효율적으로 유지하는데 계약적 지배구조보다 관계적 지배구조가 더 효율적이라는 점을 지적하고 있다. 특히 관계적 지배구조와 성과와의 긍정적 관계는 체인 내 가맹점 간 경쟁수준이 낮은 경우에 더욱 강화되어진다.

연구결과는 다음과 같이 이론적 그리고 실무적으로 시사점을 제공하고 있다. 이론적 측면에서 보면, 본 연구결과는 가맹본부와 가맹점 간의 상호신뢰와 몰입이 성과에 긍정적으로 영향을 미친다는 다른 선행연구결과와 일치하고 있다(Zaheer and Venkatraman 1995: Baker, Gibbons and Murphy 2002, 조규호, 전달영 2003).

분석결과, 프랜차이즈 조직에서 관계적 지배구조의 효과가 본부와 가맹점 간 협력을 강화해주고, 갈등해소 뿐 만 아니라 학습과 지식을 전달하는데 계약적 메커니즘보다 훨씬 더 효과적이라는 것을 보여주고 있다(Gulati, 1995; Poppo, et al., 2002). 프랜차이즈 조직에서 두 파트너간의 신뢰가 계약비용을 절약하게 하고, 감시필요성을 줄이며, 오히려 기본적인 계약에 잘 순응하도록 도움을 줄 수 있다는 것이다.

실무적 측면에서 보면 프랜차이즈 계약은 기본적 파트너십의 기본적인 역할을 하지만, 궁극적으로 가맹점의 성과를 높여주는 것은 관계적 지배구조이다.

이 관점을 더 자세히 들여다보면, 첫째, 본부와 가맹점간의 상호신뢰는 기회주의를 줄이거

나 제거함으로서 거래와 협상비용을 줄여준다. 반면, 명문화된 공식계약은 작성하고, 감시하고, 집행하는데 과도한 비용과 시간이 필요하다. 그래서 프랜차이즈 본부와 가맹점 간 장기적 관계는 계약적 안전장치(contractual safeguard)로서 관계적 지배구조를 활용하는 것이 바람직하다.

둘째, 프랜차이즈에서 계약의 남용은 타협할 수 없는 갈등과 궁극적으로 가맹점의 성과에 치명적인 역기능적 결과를 유발할 수 있다. 심지어 명문화된 계약이 존재한다하더라도 분쟁발생시 법적 제재를 강구하기가 어렵다. 기존 프랜차이즈에 대한 선행연구들이 가맹점을 관리하기위해 이용하는 중요한 수단이 계약파기라고 주장한 학자도 있으나(Rubin 1978, Brickley and Dark 1987, Strutton, Pelton and Lumpkin, 1995) 지적한대로 그런 제재를 선택하는 것이 그에 따른 비용과 바람직하지 않은 결과가 너무 크다는 것이다.

셋째, 신뢰는 가맹본부와 가맹점 간 학습과 지식교류를 용이하게 해준다. 이것은 파트너간 강력한 상호관계를 구축해준다. 신뢰를 쌓고 있는 파트너는 기회주의 행동에 대한 두려움에서 해방될 수 있다. 개방과 투명성은 지적교류를 원활하게 해준다. Bradach(1997)에 따르면 프랜차이즈 본부와 가맹점 간 상호신뢰는 가맹점의 지역시장에 대한 지식과 사업경험을 통하여 전체 시스템의 가치를 향상시킬 수 있으며 가맹점으로 부터 훌륭한 의견을 수용함으로서 수익을 창출하게 하는 지역적응성(local response)을 강화시킬 수 있다.

넷째, 가맹점간 내부 경쟁은 본부에 대한 신뢰를 무너뜨리고 결국 성과에도 부정적이다. 시장 확장을 위해 지역 내 가맹점을 추가할 경우 기존 가맹점에 대한 이해와 설득이 무엇보다 중요하다. 가맹점 간 내부경쟁이 치열하면 결국 본부에 대한 신뢰가 떨어지고 또한 계약에 대해서도 불신으로 이어질 수 있다. 이 경우 오히려 상세하게 합의된 계약이 가맹점을 설득시키는데 유효할 수 있다.

다섯째, 본부와 가맹점 간의 관계에서 계약적 지배구조가 성과에 기여하지 않는다 하더라도 그 역할에 대하여 완전히 부정하는 것은 아니다(예를 들면, 기업성과의 영향력 β=.121, $p<.10$). 오히려 공식적인 프랜차이즈 계약이 쌍방의 기본적 요구사항을 정하고 상호관계를 유지하기 위한 기본적 틀로서 역할을 할 수도 있다. 계약은 본부와 가맹점간 이행해야 될 중요한 영역을 미리 정하여 공정한 원칙에 따라 합리적으로 운영할 수 있도록 함으로써 장기간 신뢰관계에 도움을 줄 수 있다.

마지막으로 다음과 같은 본 연구의 한계가 있을 수 있다. 첫째, 본 연구에 참여한 가맹점주나 점장들의 운영경험이 평균1년5개월이다. 본부의 조언에 따르면 신규가맹점이 개업 후 2년은 지나야 수익이 발생한다고 하는데 짧은 경력을 지닌 가맹점들의 성과를 답하는데 무리가 있을 수 도 있다. 향후 연구에서는 어느 정도 경력을 고려하여 조사하는 것이 정확한

결과를 예측할 수 있을 것이다. 두 번째는 가맹점에게 계약적 지배구조 설문문항을 평가하도록 한 점이 문제가 될 수 있다. 과연 본인들한테 구속적인 질문에 대해 얼마나 솔직하게 대답할 것인가? 이러한 점이 결과를 왜곡할 수 도 있다. 향후 연구에서는 본부의 의견과 가맹점의 의견을 비교하는 것도 필요하다고 본다.

참고문헌

[1] 김응수·임영균 (2006). 프랜차이즈 시스템 내 가맹본부의 생존요인분석,『경영학연구』, 35-5, 1589-1614.

[2] 김천서·김의근·전재균 (2004). 레밀리 레스토랑 고객의 서비스회복 공정성지각과 신뢰 및 행동의도간의 인과관계 연구,『관광레저연구』, 16-3.

[3] 조규호·전달영 (2003). 프랜차이즈 시스템에서 운영구조와 관계특성이 신뢰 및 몰입에 미치는 영향,『경영학연구』, 32-5. 1265-1289.

[4] 박광서·안종석 (2000).국제합작투자기업의 지배구조가 관계지속을 위한 특유적 자산의 투자 및 환경 불확실성 지각에 미치는 영향.『경영학연구』, 29-2. 109-131.

[5] 양신철·한경수·김영국 (2005). 외식프랜차이즈의 관계적 특성과 관계의 질이 장기지향성에 미치는 영향,『관광연구저널』, 19-3

[6] 오세조·김상덕·오일두 (2003).관계기간에 따른 통제기제 및 관료화가 프랜차이즈 가맹점의 결속과 관계만족에 미치는 영향.『유통연구』, 8-1.

[7] 설훈구·강성욱·박기용 (2007). 레밀리, 패스트 푸드 레스토랑의 선택속성을 통한 시장세분화에 관한 연구,『관광레저연구』, 19-4.

[8] Agrawal, D., and R. Lal (1995). "Contractual Arrangements In Franchising: An Empirical Investigation", *Journal of Marketing Research,* 32(3), 213-221.

[9] Artz, K. and Brush, T. (2000). "Asset specificity, Uncertainty, and Relational norms". *Journal of Economic Behavior and Organization*, 41, 337-362.

[10] Baker, G., R. Gibbons, and K. J. Murphy (2002). "Relational Contracts and the Theory of the Firm," *Quarterly Journal of Economics,* 117(1), 39-84.

[11] Bradach, J. L. (1997). "Using the Plural Form in the Management of Restaurant Chains," *Administrative Science Quarterly,* 42(2), 276-303.

[12] Bradach, J. L. (1997). *Franchise Organization*, Harvard Business School Press.

[13] Bradach, J. L., and R. G. Eccles (1989). "Price, Authority, and Trust: From Ideal Types to Plural Forms," *Annual Review of Sociology,* 15, 97-118.

[14] Brickley, J.,, and F. H. Dark (1987). "The Choice of Organizational Form: The Case of Franchising," *Journal of Financial Economics,* 18(2), 401-420.

[15] Cochet, O., Dorman, J. and Ehrmann, T.(2008)" Capitalizing on Franchisee Autonomy: Relational Forms of Governance as controls in Idiosysncratic

Franchise Dyads," *Journal of Small Business Management* ,46(1), 50-67.

[16] Das, T. K., and B. S. Teng (1998). "Between Trust and Control: Developing Confidence in Partner Cooperation in Alliances," *Academy of Management Review,* 23(3), 491-512.

[17] Gulati, R. (1995). "Does familiarity breed trust?. The implications of repeated ties for contractual choice in alliance". *Academy of Management,* 38. 85-112.

[18] Hair, J. F., R. E. Anderson, R. L. Tatham, and W. C. Black (1998). Multivariate Data Analysis. *Upper Saddle River*, NJ: Prentice-Hall

[19] Heide, J. B. (1994). "Interorganizational Governance in Marketing Channels," *Journal of Marketing,* 58(1), 71-85.

[20] Hill, C. (1990). "Cooperation, Opportunism, and the Invisible Hand; Implication for Transaction Cost Theory". *Academy of Management Review,* 15. 500-513.

[21] Ivens, B. S. and K. J. Blois (2004). "Relational Exchange Norms in Marketing: A Critical Review of Macneil's Contribution," *Marketing Theory,* 4(3), 239-263.

[22] Jones, C., W. S. Hesterley, and S. P. Borgati (1997). "A General Theory of Network Governance: Exchange Conditions and Social Mechanisms," *Academy of Management Review,* 22(4), 911-945.

[23] Kaufmann, P. J., and S. Eroglu (1999). "Standardization and Adaptation in Business Format Franchising," *Journal of Business Venturing,* 14(1), 69-85.

[24] Kaufmann, P. J., and L. W. Stern (1998). "Relational Exchange Norms, Perceptions of Unfairness, and Retained Hostility in Commercial Litigation," *Journal of Conflict Resolution,* 32(3), 534-552.

[25] Klein, B. (1995). "The Economics of Franchise Contracts," *Journal of Corporate Finance,* 2(1/2), 9-37.

[26] Klein, B. and K. Murphy (1998). "Vertical Restraints as Contract Enforcement Mechanisms," *Journal of Law and Economics,* 31(2), 265-297.

[27] Macneil. I. R. (1980). The New Social Contract: An Inquiry into Modern Contractual Relations. *New Haven*, CT: Yale University Press.

[28] Ness H. and Haugland S.A. (2005)."the evolution of governance mechanism and negotiation strategies in fixed-duration interfirm relationship". *Journal of Business Research,* 58, 181-191.

[29] Pazanti, I., and M. Lerner (2003). "Examining Control and Autonomy in the Franchisor-Franchisee Relationship," *International Small Business Journal,* 21(2), 131-159.

[30] Poppo, L., and T. Zenger (2002). "Do Formal Contracts and Relational Governance Function as Substitutes or Complements?", *Strategic Management Journal,* 23(8), 707-725.

[31] Roath, A. S., Miller, S.R., and Cavusgil, S.T.(2002). "A conceptual Framework of Relational Governance in Foreign Distribitor Relationship," *International Business Review,* 11, 1-16.

[32] Rubin, P. H. (1978). "The Theory of Firm and the Structure of the Franchise Contract," *Journal of Law and Economics*, 21(1),

[33] Shane, S. A.(1996). "Hybrid Organizational Arrangements and Their Implications for Firm Growth and Survival; A Study of New Franchisors," *Academy of Management Journal,* 39(8), 216-34.

[34] Stassen, R. E., and R. A. Mittelstaedt (1995). "Territory Encroachment In Maturing Franchise Systems," *in Franchising: Contemporary Issues and Research*. Eds. P. J. Kaufmann and R. P. Dant. New York: The Haworth Press, 27-48.

[35] Strutton, D., L. E. Pelton, and J. R. Lumpkin (1995). "Psychological Climate in Franchising System Channels and Franchisor-Franchisee Solidarity, *Journal of Business Research,* 34(2), 81-91.

[36] Williamson, O.(1985). *The Mechanism of Market Governance*. NY. Oxford Press.(1981). "The economics of organization; the transaction cost approach," *American Sociology*, 87(3), 548-577.

[37] Wu, F. and Cavusgil, S.T. (2006). "Organizational Learning, Committment and Joint value Creation in interfirm Relationship, *Journal of Business Research,* 59(2), 81-91.

[38] Lee, Yikuan and Cavusgil, S.T. (2006). "Enhancing alliance performance; the effects of contractual-based versus relational-based governance," *Journal of Business Research,* 34(2), 59. 896-905.

[39] Uzzi, B. (1997). "Social structure and compeittion in intrfirm networks; the

paradox of embeddedness," *Administrative Science Quarterly,* 39, 35-67.

[40] Zaheer, A., B. McEvily, and V. Perrone (1998). "Does Trust Matter? Exploring the Effects of Interorganizational and interpersonal trust on performance," *Organization Science.* 9(2), 141-159.

[41] Zaheer, A., and N. Venkatraman (1995). "Relational Governance as an Interorganizational Strategy: An Empirical Test of the Role of Trust in Economic Exchange," *Strategic Management Journal,* 16(5), 373-392.

[투고일: 2009. 12. 26] [심사(수정)일: 2009. 2.10] [게재확정일: 2009. 2.15]