Journal of Convergence for Information Technology
Vol. 12. No. 5, pp. 100-106, 2022
e-ISSN 2586-4440
DOI : https://doi.org/10.22156/CS4SMB.2022.12.05.100

자기조직화지도 클러스터링을 이용한 종단자료의 탐색적 분석방법론

조영빈
건국대학교 국제비즈니스학부 경영학과 교수

An Exploratory Methodology for Longitudinal Data Analysis Using SOM Clustering

Yeong Bin Cho
Professor, Division of International Business Dept. of Business Administration

요 약 종단연구는 동일 대상에 대하여 반복적으로 측정한 종단자료를 기반으로 하는 연구방법을 말한다. 대부분의 종단분석 방법은 예측이나 추론에 적합하고, 탐색적 목적으로 사용하기에는 적합하지 않은 경우가 많다. 본 연구에서는 종단자료를 분석하는 탐색적 방법을 제시한다. 이 방법은 자기조직화지도기법을 사용하여 종단자료를 군집화 하여 최선의 군집 수를 정한 후 종단궤적을 찾는 방법이다. 제안한 방법론은 고용정보원의 종단자료에 적용되었으며, 총 2,610개의 샘플에 대하여 분석을 하였다. 방법론을 적용한 결과 패널 별로 시계열적으로 군집 화되는 결과를 얻었다. 이는 종단자료를 사전에 클러스터링하고 다층 종단분석을 하는 것이 더욱 효과적이라는 사실을 나타낸다.

주제어 : 종단자료, 종단분석, 잠재성장모형, 자기조직화지도기법, 종단궤적

Abstract A longitudinal study refers to a research method based on longitudinal data repeatedly measured on the same object. Most of the longitudinal analysis methods are suitable for prediction or inference, and are often not suitable for use in exploratory study. In this study, an exploratory method to analyze longitudinal data is presented, which is to find the longitudinal trajectory after determining the best number of clusters by clustering longitudinal data using self-organizing map technique. The proposed methodology was applied to the longitudinal data of the Employment Information Service, and a total of 2,610 samples were analyzed. As a result of applying the methodology to the actual data applied, time-series clustering results were obtained for each panel. This indicates that it is more effective to cluster longitudinal data in advance and perform multilevel longitudinal analysis.

Key Words : Longitudinal Data, Longitudinal Analysis, Latent Growth Model, Self-Organizing Model, Longitudinal Trajectory.

1. 서론

빅 데이터 시대가 되면서 수많은 데이터가 축적되고 있다. 청소년, 교육, 노동, 아동, 노인 등의 분야에서 계층별, 기능별로 정기적인 종단 조사가 시행되면서 종단자료도 점차 많아지고 있다[1]. 또한 비즈니스 분야의 종단자료 규모도 커지고 있다. 모바일 결제, 신용카드가 일반화되면서 개인별 구매내역이 장기간 저장되고, 인터넷, 모바일 로그기록도 유지되면서 비즈니스 분야의 종단자료도 확충되고 있다[1].

종단 자료를 바탕으로 하는 종단 분석은 시간흐름에 따른 반응의 변화를 추정하고 변화량과 개인 내 변화

*Corresponding Author : Yeong Bin Cho(ybcho111@kku.ac.kr)
Received March 13, 2022
Revised April 4, 2022
Accepted May 20, 2022
Published May 28, 2022

(within-individual change)를 추정한다[2,3]. 대표적인 종단분석방법인 잠재성장모형은 ① 시간흐름에 따른 반응변화를 추정하기 위하여 비조건 모델을 휴리스틱 방법으로 찾아서 적합도가 높은 모델을 찾고 ② 개인 내 변화와 영향변수를 찾기 위하여 조건 모델을 만들어 종단 인과관계 검증을 하게 된다[4]. 다시 말해서 시간의 흐름에 따라 변화하는 종단 궤적(longitudinal trajectory)을 찾아서 검증하는 과정이 핵심이라 할 수 있다[5].

본 연구에서는 자기조직화지도(SOM: Self-Organizing Map) 클러스터링 기법을 사용하여[6], 종단자료의 잠재 종단궤적을 찾는 방법론을 제안한다. 이를 위하여 자기조직화지도기법을 사용하여 패널 자료를 클러스터링하여 클러스터번호를 배정하고, 이를 시간의 흐름에 따라 재분류하여 클러스터 번호들의 순서 자료로 변환한다[1]. 이렇게 하면 패널 자료는 부여된 클러스터 번호 순서로 변환되어 나타내게 된다. 클러스터 번호의 궤적은 패널의 종단 궤적으로 간주할 수 있을 것이다. 그 이후 종단 궤적 중 타당성을 갖고 있는 궤적에 대한 종단분석방법에서의 궤적 검증을 실시하게 된다. 이러한 과정은 본질적으로 종단 궤적을 찾는 탐색적(exploratory) 연구가 된다.

이러한 방법론을 실현하기 위해서는 최선의 클러스터 수를 결정해야 한다. 클러스터 수를 정하는 평가지표는 다양하게 개발되어 사용되고 있지만, 데이터 집합의 특성이나 클러스터링 방법에 상관없이 모든 상황에서 최적 클러스터 수를 결정하는 평가지표는 아직 없는 것으로 알려져 있다 [7-9]. 본 연구에서는 보편적인 평가지표인 실루엣(Silhouette)지표와 칼린스키와 하라바즈(Calinski & Harabasz)지표를 순차적으로 사용하여 최선의 클러스터 수를 정하였다. 제안한 방법론의 타당성을 검증하기 위하여 한국고용정보원의 고령화연구패널(KLoSA)에 적용하였다.

본 논문은 다음과 같이 구성된다. 2장에서는 본 연구에서 제안한 방법론의 절차와 내용을 설명한다. 3장에서는 제안한 방법론을 실제 데이터에 적용하고 실제 종단 궤적을 추출하고 적용하는 방법을 제시한다. 4장에서 결론과 향후 연구방향 및 내용을 기술한다.

2. 제안 방법론

본 연구의 방법론은 Fig. 1에서와 같이 크게 세 가지 단계로 구분할 수 있다. 1. 종단자료 전처리; 2. 자기조직화지도 클러스터링; 3. 최선 클러스터 궤적 도출 등이다. 종단자료 전처리는 시간 흐름에 따른 결측치(missing value)와 이상치(outlier)를 처리하고 클러스터링을 위한 데이터 집합을 만드는 단계이다.

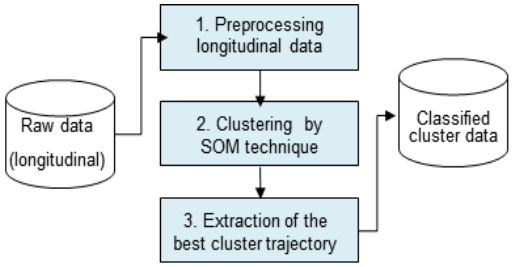

Fig. 1. Research Process

두 번째 자기조직화지도 클러스터링은 기계학습 방법 중 대표적인 비지도 학습방법인 자기조직화 지도(SOM: Self Organizing Map) 기법을 이용하여 클러스터링을 시행하는 단계이다. 세 번째 단계는 최선 클러스터 도출로 가장 성능이 좋은 클러스터 수를 결정하고 최선 클러스터에 따른 종단자료의 궤적을 결정한다. 이후 결정된 클러스터의 궤적에 따라 주어진 종단자료를 구분함으로써 향후 추론과 예측을 위한 데이터 클러스터를 제공한다.

2.1 종단자료 전처리

종단자료 전처리는 종단자료 분석과 클러스터링을 위한 준비단계이다. 종단자료는 조사 특성상 결측치가 발생할 가능성이 크다. 왜냐하면 동일 패널에 대해서 반복적으로 조사를 하는데 특정 조사시점에 응답을 하지 않으면 결측치가 발생하게 된다. 종단자료의 결측치와 이상치에 대한 전처리(pre-processing)는 반드시 필요하다[5].

n명의 패널을 대상으로 m개의 항목을 시간 흐름에 따라 T번 반복 조사했다고 가정하자. 그러면 인스턴스 X_{ijt}는 i 패널 $(i=1,2,...,n)$에 대한 j 항목 $(j=1,2,...,m)$의 t번째 $(t=1,2,...,T)$의 측정치로 정의할 수 있다. 각각의 i 패널에 대해서 패널 ID를 제외하면 $X_{ij\cdot}=(x_{i1t}, x_{i2t}, \cdots, x_{imt})^t$로 표현할 수 있다. 데이터 집합은 3차원 형태로 구성할 수 있고 이런 형태의 종단자료를 롱타입(long type)이라 한다. 또 다른 형태의 종단자료는 와이드 타입(wide type)으로 다음과 같은 형태로 표현

$X_{i\,\cdot\,\cdot} = (x_{i11}, x_{i12}, \cdots, x_{i1T}, x_{i21}, x_{i22}, \cdots, x_{i2T}, \cdots, x_{im1}, x_{im2}, \cdots, x_{imT})^t$ 할 수 있다. 이는 i 패널을 기준으로 j 조사항목의 t 측정값을 T개까지 배열하고, 다음번 j 조사항목의 측정값으로 배열하는 방식이다. 와이드 타입은 2차원 형태로 표현된다. 롱타입은 종단자료를 개개의 데이터로 처리할 수 있는 반면, 와이드 타입은 패널 별로 시간흐름에 따른 변화를 반영한 순서(sequence) 데이터로 인식한다는 점이 다르다고 볼 수 있다. 물론 롱타입도 데이터 처리를 해서 와이드 타입으로 변환할 수 있다. 종단분석은 시간흐름에 따른 체계적인 변화가 있는지 파악하고, 이러한 체계적인 변화가 개인별로 차이가 있는지를 분석한다. 따라서 대부분의 종단분석방법은 패널 별로 항목의 순서(sequence)의 변화를 감지하는 와이드 타입을 많이 사용한다. 그렇지만 본 연구에서는 패널 별로 항목의 순서 변화를 감안하지 않고 롱 타입처럼 종단데이터를 개개의 데이터로 인식하여 클러스터링 한 후 패널 별 순서변화 궤적을 추적하는 방법을 사용한다.

2.2 자기조직화지도 클러스터링

클러스터링은 전형적인 비지도 학습(unsupervised learning)기법이다. 본 연구에서는 자기조직화 지도(SOM: Self Organizing Map) 기법을 이용하여 클러스터링 한다. 자기조직화지도기법은 인공신경망 기반의 한 방법으로 고차원 데이터를 투영(mapping)하여 일반적으로 2차원 그리드로 표현한다. 투영은 여러 기하학적 관계와 토폴로지 및 메트릭 관계를 대략적으로 유지하기 때문에 원래 데이터의 관계를 반영하여 클러스터 한다고 볼 수 있다[6]. 자기조직화지도기법을 이용하면 X축과 Y축의 값을 변화시키면서 2차원으로 클러스터링을 할 수 있어 데이터 시각화에 효과적인 기법으로 알려져 있다[6]. 자기조직화지도기법은 원 자료의 확률밀도함수를 그대로 따르고, 설명가능하고 간단하면서 시각적으로 출력된다. Fig. 2와 같이 시각적으로 출력 될 수 있기 때문에 종단궤적을 시각화 할 수 있다는 장점이 있다. 롱타입의 데이터를 입력 자료로 사전에 정한 개수에 따라 자기조직화지도 클러스터를 생성한다. 가로 p개, 세로 q개로 $p \times q$ 개의 자기조직화지도 클러스터를 형성한다고 하면, 클러스터의 집합 $C = \{C_{11}, C_{12}, ..., C_{pq}\}$을 이루게 된다. 종단자료의 인스턴스들이 다음과 같이 배정될 수 있다. $(x_{i1t}, x_{i2t}, \cdots, x_{imt})^t \rightarrow C_{kj} \quad k = 1,2,...,p,$ $j = 1,2,...q.$ 이렇게 배정된 개개의 인스턴스를 패널 별로 분류하고 측정시점 순서대로 다시 분류하면 클러스터별 궤적으로 변환할 수 있다. 각 클러스터의 특성을 규명하면 패널 궤적변화에 관한 의미를 알 수 있다. 이러한 궤적은 대표적인 종단분석방법인 잠재성장모형의 비조건모형(unconditional model)에서 반응변수의 시간흐름에 따른 궤적을 보고 체계적인 경향성을 검증하는 절차와 유사하다. 하지만 경향성을 파악하는 데 있어서 반응변수만 한정하지 않고, 모든 조사항목의 대체적인 경향성을 보기 위하여 궤적을 찾는다는 점이 다르다.

Fig. 2. Longitudinal Trajectory

2.3 적정 클러스터 궤적 도출

일반적으로 모든 클러스터링 기법은 최적 클러스터 수를 정하기 어렵다. 자기조직화지도기법도 사전에 클러스터 개수를 정할 수 있으나, 몇 개의 클러스터가 가장 최적인지는 알 수 없다. 따라서 클러스터의 수의 적정성을 평가할 지표가 필요하다. 클러스터의 개수를 정하기 위해서 클러스터링 적합성 평가 지표는 다양하게 개발되어 있다[7]. 본 연구에서는 실루엣(Silhouette)지표와 칼린스키와 하라바즈(Calinski & Harabasz) 지표를 순차적으로 사용하여 최선 클러스터 수를 찾는다.

먼저 실루엣(Silhouette) 지표는 특정 데이터가 특정 클러스터로 분류될 경우 해당 특정 데이터가 가장 인접한 클러스터가 아닌 해당 클러스터로 분류된 정당성 정도를 측정한다. 특정 데이터 D의 실루엣 통계량은 다음과 같이 정의할 수 있다[10]. Sil(D)=(b(D)-a(D))/ Max [a(D),b(D)] 여기서 a(D)는 같은 클러스터 내 다른 데이터들로부터 특정 데이터 D가 얼마나 떨어져 있는지를 나타내는 이격성(remoteness) 정도를 나타낸다. b(D)는 특정 데이터 D와 가장 가까운 다른 클러스터의 특정 데이터로부터 해당 특정 데이터 D가 떨어져 있는 이격

성 정도를 나타낸다. 이는 이격성 b(D)는 대상 특정 데이터와 모든 외부 클러스터들의 특정 데이터들사이에서 계산된다는 것을 의미한다. k개 클러스터가 있다고 가정할 때 계산된 k-1개의 이격성 중 가장 작은 값이 공식에서 취한 b(D)가 된다. 실루엣 통계량은 -1(이론적으로 특정 데이터에 대한 최악의 클러스터링)에서 +1(이론적으로 특정 데이터에 대한 최고의 클러스터링)까지 변동할 수 있다. 실루엣 지표는 모든 데이터의 실루엣 값에 대한 산술평균을 사용하고 값이 클수록 클러스터링 품질이 더 좋은 것으로 알려져 있다. 실루엣 지표는 클러스터링의 품질을 측정할 수 있는 보편적인 지표지만 클러스터별로 동일 수치가 나올 가능성이 높아서, 최선 클러스터 숫자를 찾기 위해서는 동일 수치(tie)을 해결할 별도 지표가 필요하다. 이를 위하여 칼린스키와 하라바즈(Calinski & Harabasz) 지표 [7,8]를 사용하였다. 칼린스키와 하라바즈(Calinski & Harabasz) 지표는 $\frac{BGSS}{WGSS} \times \frac{N-k}{k-1}$로 표시된다. 여기서 WGSS는 그룹 내 분산(Within Group Sum of Square), BGSS는 그룹 간 분산 (Between Group Sum of Square), N을 총 데이터 개수, k를 클러스터 수이다. 수식 값이 최대가 되는 k를 클러스터 개수로 선택한다 [11]. 칼린스키와 하라바즈(Calinski & Harabasz) 지표는 소규모 샘플, 변수, 클러스터 수에서 다른 클러스터링 지표에 비하여 우수한 성능을 보였다[8]. 그렇지만 일반적인 조건에서 다른 지표보다 항상 우수한 성능을 보이지는 못했다. 따라서 본 연구에서는 실루엣 지표와 칼린스키와 하라바즈 지표를 순차적으로 사용하여 최선 클러스터 수를 정한다. 그 이후 종단 데이터는 ID별로 재분류되고 개별 종단 데이터는 시간흐름에 따른 클러스터들 궤적으로 표시될 수 있다. 이는 바로 해당 패널의 동적 궤적이 된다. 이러한 원 데이터의 동적 궤적을 기준으로 종단 데이터를 그룹화 할 수 있고 그룹화된 종단 데이터들은 유사한 궤적을 가진 하부 그룹이라 할 수 있다. 이러한 방법으로 그룹화 된 종단 데이터는 그룹화 되지 않은 원래의 종단 데이터에 비해서 그룹 내 변동성은 작고 그룹 간 변동성이 큰 데이터 집합이 된다.

3. 실제 데이터에 적용

3.1 적용 데이터 및 기술통계량

본 연구의 방법론을 적용한 자료는 한국고용정보원의 고령화연구패널(KLoSA)이다. 이 자료는 한국고용정보원이 2006년부터 2018년까지 격년으로 총 7차에 걸쳐서 조사하였다. 고령화연구패널의 개요는 제주도이외 지역 거주민 중 45세 이상 중·고령자를 대상으로 조사했고, 짝수 연도에 패널을 대상으로 조사를 실시하고 있다. 본 연구에서는 1차(2006년)부터 6차(2016년)까지 계속적으로 임금근로자로 근무한 대상자로 한정하였다. 총 대상자는 435명이었고, 6년간 조사를 합산하면 총 샘플 수는 2,610개였다. 분석을 위하여 사용한 소프트웨어는 SPSS statistics 27과 SPSS Modeler였다.

사용한 데이터는 직무만족도 관련 데이터로 변수는 임금수준, 임금만족도, 고용안정성, 전반적 만족도, 연령, 학력, 성별 등 7개였다. 임금수준은 연간 임금총액을 자연로그로 치환한 뒤 표준화하여 사용하였다. 임금만족도, 고용 안정성, 전반적 만족도는 1점(매우 그렇다)~4점(전혀 그렇지 않다)으로 4점 척도를 사용하였다. 3가지 변수는 탐색적 요인분석을 사용하여 하나의 변수로 변환하였으며, 크롬바 알파 값은 0.79였다. 연령도 표준화하여 사용하였다. 학력은 그대로 사용하였다. 성별은 여성 5, 남성 1로 조사되어 있어서 여성을 0으로 치환하여 코딩하였다. 사용한 변수들의 기술통계는 Table 1과 같다.

데이터 집합은 패널 435명에 대해 5개의 조사항목을 6개 기간 동안 수집한 자료이다. 결측치와 이상치는 사전에 제거하였고, 롱타입(long type)으로 정리하였기 때문에 2,610개의 데이터를 자기조직화지도기법의 클러스터링을 위한 입력 자료로 사용하였다.

Table 1 Descriptive statistics of attributes

		Min.	Max.	Avg.	SD
Satis-faction	2006	-2.728	2.591	.0000	.999
	2008	-2.770	2.488	.0000	.999
	2010	-3.039	2.599	.0000	.999
	2012	-3.079	2.552	.0000	.999
	2014	-3.661	2.917	.0000	.999
	2016	-3.772	2.819	.0000	.999
Salary	2006	-3.459	2.247	.0000	.999
	2008	-3.913	2.224	.0000	.999
	2010	-3.851	2.489	.0000	.999
	2012	-3.751	2.359	.0000	.999
	2014	-3.883	2.821	.0000	.999
	2016	-4.677	2.554	.0000	.999
Age		-1.208	3.091	.0000	.998
Education		1	4	2.61	.996
Gender		0	1	.68	.468

3.2 자기조직화지도 클러스터링 및 최선 클러스터 결정

전술한 바와 같이 자기조직화지도기법을 이용하여 클러스터링하기 위하여 SPSS Modeler의 SOM 모듈을 사용하였다. 클러스터 수는 p와 q값을 1~9까지 변화시켜 총 81개의 클러스터를 생성하였다. 분석 데이터 집합의 조사기간이 6개 기간이므로 $p \times q$는 최소 6 이상이어야 궤적을 확인할 수 있다. 81개의 생성된 클러스터 중 최선 클러스터를 결정하기 위하여 모든 클러스터에 대하여 실루엣 지표를 계산하였다. Table 2에 제시한 바와 같이 실루엣 지표 값은 0.3에서 0.5의 분포를 보였다. 실루엣 값이 큰 0.5인 클러스터는 2×3 등 12개 클러스터였다. 2×3, 1×3, 3×2 등 3개의 클러스터는 실제 클러스터 숫자는 본 연구의 종단 측정 횟수인 6보다 적기 때문에 제외하였다.

Table 2. Values of Silhouette Index for the SOM Clusters

Silhouette Index	SOM Cluster
0.5	2×3*, 1×3*, 3×2*, 5×3, 6×3, 4×4, 5×4, 6×4, 3×5, 4×5, 4×6, 3×7
0.4	3×1,…7×1, 9×1, 4×2,…9×2, 3×3, 4×3, 7×3, 8×3, 9×3, 1×4, 3×4, 7×4, 8×4, 9×4, 1×5, 2×5, 5×5,…8×5, 1×6, 2×6, 3×6, 5×6,…9×6, 1×7, 3×7, 6×7
0.3	8×1, 2×4, 9×5, 2×7, 5×7, 7×7, 8×7, 9×7

* : # of clusters is smaller than the # of the longitudinal waves.

실루엣 지표가 0.5인 9개 클러스터 중 최선 클러스터를 선정하기 위하여 칼린스키와 하라바즈 지표를 계산하였다. Table 3에는 칼린스키와 하라바즈지표 값을 제시하였다. 가장 지표 값이 큰 클러스터는 4×4 클러스터였다. 4×4 클러스터에서 실제 생성된 클러스터 수는 9개 클래스로 나타났다.

Table 3. values of Calinski & Harabasz Index

SOM	# of Clusters	Calinski & Harabasz
5×3	8	165.31
6×3	11	137.54
4×4	9	170.83*
5×4	8	121.58
6×4	10	114.47
3×5	8	162.96
4×5	7	114.96
4×6	9	102.81
3×7	8	115.45

* : the highest value

3.3. 클러스터링 결과

Fig. 3에 제시한 것과 같이 4×4 클러스터의 9개 클래스 중 X=3, Y=0가 32.4%로 가장 많았고, X=0, Y=0 15.6%, X=0, Y=2 12.2%의 순이었다.

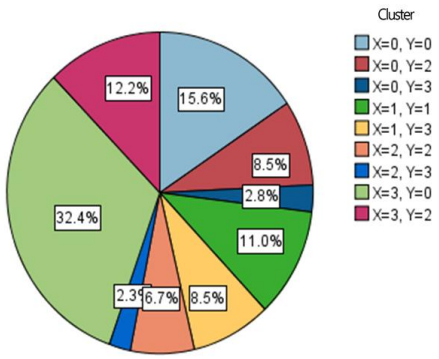

Fig. 3. The Pie Chart of 4×4 cluster

패널 궤적을 구하기 위하여 패널 ID와 시간에 따라 다시 분류하여 Table 4에 제시하였다. 435개 패널 중 426개 패널은 시간의 흐름에 따라 같은 클래스로 분류되었고, 9개만이 달라서, 전체 패널의 97.6%가 시간에 따라 같은 클래스에 속하는 궤적을 보였다. 다른 클래스에 속한 9개 패널은 03클래스와 13클래스만을 옮겨 다녔다. 결과를 종합하면 종단자료를 클러스터링을 하면 시간흐름에 상관없이 그룹화가 가능하다는 것을 의미한다.

Table 4. class×year cross table at 4×4 cluster

class	2006	2008	2010	2012	2014	2016	sum
00	68	68	68	68	68	68	408
02	37	37	37	37	37	37	222
03	10	11	12	13	13	14	73
11	48	48	48	48	48	48	288
13	39	38	37	36	36	35	221
22	29	29	29	29	29	29	174
23	10	10	10	10	10	10	60
30	141	141	141	141	141	141	846
32	53	53	53	53	53	53	318
sum	435	435	435	435	435	435	2,610

4×4 클러스터의 9개 클래스는 각각 그룹화가 가능한 성격을 갖고 있었다. Fig. 4에 제시한 바와 같이 평균 연봉 변수는 클래스 간 상호 교차가 거의 없었다. 다

른 변수도 비슷한 모습을 보였다. 이는 9개의 클래스를 다른 데이터 집합으로 구분하는 것이 타당하다는 것을 나타낸다. 따라서 본 연구의 대상인 데이터 집합은 9개의 데이터 집합으로 구분하여 추론 및 예측하는 다층모형을 적용하는 것이 더 좋다고 볼 수 있다.

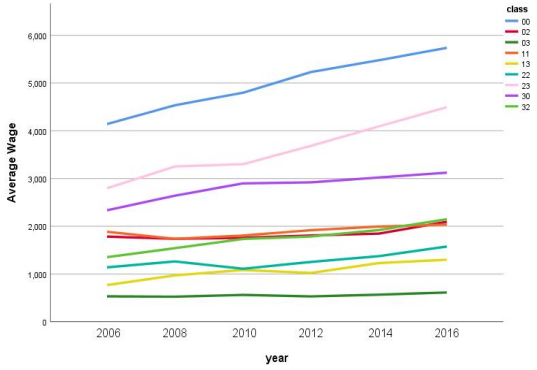

Fig. 4. Serial changes of Average Wage of each cluster

4. 토의 및 결론

본 연구에서는 자기조직화지도 클러스터링 기반의 탐색적 종단자료 분석방법론을 제시하였다. 제시한 방법론을 실제 자료에 적용하여 의미 있는 종단 궤적을 추출했다. 이렇게 추출된 종단궤적은 잠재성장모형과 같은 종단분석기법에 사용할 수 있을 것으로 예상하였다. 그런데 예상과는 달리 전체 종단 데이터의 97.6%가 같은 클래스의 종단 궤적을 갖는 것으로 나타났다. 이 결과로 대부분의 패널이 같은 클러스터링 클래스를 갖게 되고 종단궤적은 변화가 없는 수평선이 된다. 이는 원 데이터 집합에 있는 잠재 종단 궤적을 찾아서 종단분석기법에 적용하는 것이 아니라, 원 데이터 집합을 9개의 데이터 집합으로 쪼개서 하위 데이터 집합에 각각 종단분석기법을 적용하는 것이 효과적이라는 것을 의미한다. 이는 본 연구의 중요한 의의라 할 수 있다. 그렇지만 아직은 제한된 데이터 집합의 결과이므로 일반화할 수는 없다.

이런 결과가 나오게 된 원인은 세 가지로 추정할 수 있다. 첫 번째는 종단분석은 동일집단에 속하는 패널간 유사성과 동일 패널에 속하는 반복측정치의 유사성을 검증하기 때문에[12], 동일 패널에 대한 반복 측정 자료를 와이드 타입으로 처리한다. 반면 본 연구는 동일 패널의 반복 측정 자료를 개개의 데이터로 보고 롱 타입으로 처리하여 클러스터링한 후, 클래스를 지정한 뒤 패널 별로 다시 구성하는 방법을 사용하였다. 이런 이유로 와이트 타입으로는 추출되지 않았던 현상이 나왔을 가능성이 있다. 두 번째로는 본 연구의 데이터 집합은 원천적으로 패널 간 변량보다 동일 패널 반복측정치의 변량이 작은 것은 아닌지 검증이 필요하다. 더 나아가 패널 간 변량과 패널 반복측정치 변량의 크기에 따라 종단 궤적의 모양이 달라지는 것은 아닌지 추가적인 연구가 필요하다. 세 번째로 클러스터링 기법 특성에 따른 결과로 생각해볼 수 있다. 자기조직화지도기법은 K-means와 같은 클러스터링기법과는 달리 2차원으로 맵핑하기 때문에, 종단자료가 갖고 있는 다차원성을 클러스터링 했을 가능성이 있다. 자기조직화지도기법뿐만 아니라 다른 클러스터링방법에 의한 결과와 비교할 필요가 있다.

사실 종단자료 분석은 상거래의 추천문제에서 중요한 의미를 갖는다. 예를 들어 특정 고객의 거래액이 지속적으로 변화하는 경우, 영향 요인으로 연령과 거주지와 같은 인구통계학적 변수나 소득수준과 소비액과 같은 사회경제적인 변수를 고려한다. 그렇지만 대부분 은 횡단적인 관점에서 분석한다. 하지만 시간흐름에 따라 체계적인 고객 구매패턴의 변화가 있다고 볼 수도 있을 것이다. 예를 들어 인터넷이 보급되지 않았던 2000년 이전에 태어난 고객과 2007년 아이폰 출시 이후에 태어난 고객의 구매패턴은 차이가 있을 것이고, 통신 관련기기에 대한 선호 정도도 연령별로 뚜렷한 격차가 있을 것이다. 이러한 격차를 기존 방법론으로 분석할 경우 시간흐름에 따른 체계적인 변화양상을 찾아내기는 어렵다. 그렇지만 종단분석방법을 사용할 경우 코호트별로 구매패턴을 규명할 수 있고, 과거 구매액의 변화양상에 대한 분석에도 종단적 영향을 제시할 수 있을 것이다. 더 나아가 상품추천에도 종단 특징을 반영하여 추천의 정확도 향상에 기여할 수 있을 것이다. 본 연구의 결과는 상품추천분야에서 사용될 수 있을 것이고 추가적인 연구가 필요하다.

본 연구의 한계점은 사용한 데이터 집합의 일반성을 담보하기 어렵다는 것이다. 다양한 데이터 집합에 대한 적용이 필요하다. 그리고 7차 조사가 있었지만 종단성을 확보하기 위하여 6차까지 자료만 분석한 것도 한계라 할 수 있다. 추가연구에서 종단성 검증에 대한 연구가 필요하다. 또한 최선 클러스터 수를 결정하는데 있어

실루엣지료와 칼리스키와 하라바즈 지표를 순차적으로 사용하였다. 그렇지만 여기서 산출된 클러스터 숫자가 부분 최적일 가능성을 배제할 수 없다. 모든 경우에 적용될 수 있는 최적 클러스터 숫자를 찾는 방법은 존재하지 않는 것으로 알려져 있지만, 전체 최적을 위한 방법을 찾아야 할 것이다. 또한 본 연구의 범위는 종단자료를 위한 탐색적 연구로 설정하였기 때문에 잠재성장모형과 같은 종단분석을 실시하지는 않았지만 탐색연구 이후 인과성을 검증하는 실증연구가 필요할 것이다.

REFERENCES

[1] Y. B. Cho. (2018). A Data Based Methodology for Estimating the Unconditional Model of the Latent Growth Modeling, *J. Digital Convergence, 16(6),* 85-93.
DOI : 10.14400/JDC.2018.16.6.085

[2] G. M. Fitzmaurice, N. M. Laird & J. H. Ware. (2012). *Applied Longitudinal Analysis,* 2nd ed. John Wiley & Sons; Hoboken; New Jersey.

[3] J. D. Singer & J. B. Willet. (2006). Longitudinal data analysis: Present status; future prospects. *In Presentation at the 45th Congress of the German Psychological Association, Nurnberg, Germany* (pp. 17-21).

[4] G. S. Kim. (2009). *Latent Growth Modeling and Structural Equation Model.* Hannarae Academy.

[5] C. Genolini & B. Falissard. (2011). Kml: a package to cluster longitudinal data, *Computer Methods and Programs in Biomedicine, 104, 112-121.*

[6] T. Kohonen. (1990). The Self-Organizing Map. *Proceedings of the IEEE, 78(9), 1464-1480.*

[7] G. W. Milligan & M. C. Cooper. (1985). An Examination of Procedures for Determining the Number of Clustera in a Data Set. *Psychometrika, 50(2),* 159-179.
DOI : 10.1007/BF02294245

[8] Y. Shim, J. Chung & I. Choi. (2006). A Performance Comparison of Cluster Validity Indices based on K-means Algorithm. *Asia Pacific Journal of Information Systems, 16(1),* 127-144.

[9] N. D. Teuling, S Pauws & E Heuvel (2022). Clustering of longitudinal data: A tutorial on a variety of approaches-. *arXiv preprint arXiv :2111.05469.*

[10] L. Kaufman & P. Rousseeuw. (1990). Finding groups in data: an introduction to cluster analysis. NewYork.

[11] R. B. Calinski & J. A. Harabasz. (1974). dendrite method for cluster analysis, *Communications in Statistics, 3, 1-27.*

[12] S. Hong. (2009). *Longitudinal Research Methodology Using Multilevel Model and Latent Growth Model.* (Online). https://www.kli.re.kr/klips/downloadCnfrncSjIemFile.do?iemNo=237

조 영 빈(Yeong Bin Cho) [정회원]

· 1985년 : 고려대학교 산업공학과 (공학사)
· 1988년 : 한국과학기술원 산업공학과(공학석사)
· 2005년 : 한국과학기술원 경영대학 경영공학(경영정보학 박사)
· 2006년 3월 ~ 현재 : 건국대학교 국제비즈니스학부 경영학과 교수
· 관심분야 : CRM, 데이터마이닝, 온라인고객
· E-Mail : ybcho111@kku.ac.kr