심층 군집화를 이용한 가계 소비패턴 분석

이영빈*)

<요약>

본 연구의 목적은 가계의 다양성을 반영하는 소비패턴을 찾고 그 요인을 파악하는 것이다. 소비패턴은 소비항목별 소비비중에 따라 유형화될 것이라는 가정 하에 통계청의 가계동향조사 자료 중 12개 소비지출 항목에 대한 소비비중을 기준으로 가계를 군집화하였다. 군집방법으로는 고차원 데이터의 비선형 관계를 포착할 수 있는 심층 군집화 방법 중 모델 구조가 간단하여 활용성이 우수한 N2D모형을 사용하였다.

주요 연구결과는 다음과 같다. 첫째, 가장 우수한 성능을 보인 군집 개수 9개로 가계소비를 유형화하였다. 그 결과 발견된 소비패턴은 교통비 중심형, 교육비 중심형, 여가생활(문화) 중심형, 여가생활(여행) 중심형, 가정 중심형, 복합형, 보건 중심형, 식료품 중심형, 주거 중심형이다. 둘째, 가구실태 변수 중 가구소득, 가구원 연령, 자동차 보유대수 등이 소비패턴에 중요한 영향을 미쳤다. 셋째, 발견된 소비패턴이 가계의 다양성을 충분히 반영하였음 소비 포트폴리오 차트를 통해 검증하였다.

*) 울산과학기술원(UNIST) 산업공학과 석사과정. E-mail: young@unist.ac.kr

Ⅰ. 서론

개인의 금융 활동 데이터는 수집과 유통이 쉬워지고 있어 점점 널리 쓰이고 있다. 분석가들은 매출 예측이나 개인화 마케팅 등의 여러 작업을 수행하기 위해서 개인의 다양성을 고려하여 금융활동을 모델링하려 한다. 예를 들어, 카드사는 정교한 빅데이터 분석을 통해 타겟 고객군을 명확히 하고 사용자의 소비패턴 별 수요에 맞는 서비스로 카드 상품을 구성한다. 이렇게 개인 맞춤화된 혜택을 제공할 때, 생산자는 비용은 줄이고 마케팅 효과는 극대화할 수 있다.

가장 일상적인 금융활동 중 하나인 소비활동을 분석할 때, 개인의 다양성은 인구통계학적 변수와 같은 단편적인 요소로 흔히 구분된다. 즉, 소비패턴이 개인의 특성으로 설명된다고 가정된다. 예를 들어 소비 항목들을 성별 및 연령별로 구분하는 경우, 40대 여성들의 서로 다른 소비성향은 모두 동일하다고 가정된다. 이렇게 소비패턴이 단순하게 나뉘는 이유는 방법이 간단하면서도 설명이 용이하기 때문이다. 또한 소비자 개개인을 추적해 행동 데이터를 수집하기는 어렵기 때문에 개인을 불분명하게 특정지을 수밖에 없다. 그러나 개인을 고정 불변하며 단순한 선형적인 요소로 구분할 경우 개인의 다양성에 대한 상세한 이해를 하기 어렵다. 또한 소비패턴의 대표성이 약해지고 결과를 신뢰하기 힘든 문제가 발생한다. 예를 들어, 각 소비패턴을 구성하는 소비자들 특성의 분산이 증가한다.

사실 소비패턴과 개인의 특성 사이에는 복잡하고 비선형적인 관계가 존재한다. 개인의 소비활동은 몇 가지 변수로는 나눌 수 없을 만큼 다양성이 크기 때문이다. 또한 같은 사람이더라도 소비 성향은 시간에 따라 변화하는데, 개인의 구매 환경과 이전 구매 경험이 다음 구매 행동에 영향을 주기 때문이다. 따라서 본 연구에서는 데이터 기반의 기계학습(Machine Learning) 알고리즘을 이용하여 가계를 유형화 함으로써 소비패턴을 찾는다. 가계는 경제생활 단위로서의 가족을 지칭하며, 소비의 주체로 개인을 대변한다. 누가 어떤 업종에 얼마를 썼는지에 대한 데이터를 가지고 소비패턴을 한다면 가계 또는 개인의 특성이 아닌 소비성향이 비슷한 사람들끼리 묶을 수 있다. 이를 구현하기 위해 군집화(Clustering) 알고리즘을 사용한다.

개인의 다양한 소비 행동을 고려하는 것은 특히 경제학에서 많이 연구되어 왔다. 하지만 이러한 방법들은 데이터 내의 선형 관계를 파악할 때만 적용이 가능하다. 반면에, 심층 신경망(Deep Neural Networks) 알고리즘은 찾아내고자 하는 비선형 시스템을 근사할 수 있어 비선형 관계를 포착하는 작업에서 활발하게 사용되고 있다. 그러나 일반적인 심층 학습(Deep Learning) 기반 군집화 방법들은 모델 구조가 복잡하고 사이즈가 커서 학습시키는 데 지나치게 큰 계산 성능을 요구하고 데이터에 과대적합될 가능성이 있어 활용도가 낮다. 따라서 본 연구에서는 축소된 사이즈의 심층 군집화 방법론으로 가계를 소비패턴에 따라 군집화하고, 각 군집의 특징을 분석하였다.

Ⅱ. 이론적 배경

1. 소비패턴의 정의

경제에서의 소비는 사람들이 자신의 욕구를 충족시키기 위하여 시장에서 상품이나 서비스를 필요에 따라 일상생활에 직접 이용함으로써 경제가 움직이도록 하는 가장 기본적인 경제활동을 말한다(김인철, 2004). 그리고 소비패턴이란 상호 연관된 소비항목들이 내재적으로 구조화되는 방식을 의미한다(정영숙, 2000). 즉, 가계의 소비항목들은 서로 영향을 주고받으며 이러한 항목들이 결합하여 하나의 패턴을 이루는 것이다(손상희, 1993). 예를 들어 식당에서의 소비가 증가하면 식료품 관련 소비는 감소한다. 따라서 소비패턴은 가계가 가진 예산을 여러 소비항목에 어떻게 분배할 것인지 결정하는 방식으로 볼 수 있다. 이러한 소비행동은 사는 곳, 가치관, 소비환경, 시장의 금리, 소득 수준 등에 많은 영향을 받게 되고, 여러 연구들이 이러한 영향을 분석하였다(Yocum, C., 2007).

현대 정부와 중앙은행들은 국가 재정이나 통화 정책을 수립하기 위해 소비자 지출을 조사한다. 예를 들어 미국 상무부 산하 경제분석국은 개인소비지출(Perconal Consumption Expenditure, PCE) 이라는 이름으로 소비자 지출에 대한 데이터를 정기적으로 수집하고, 노동통계국에서 매년 소비지출조사(Consumer Expenditure Survey, CES)를 실시한다. 한국에서는 통계청이 가계동향조사라는 이름으로 가계의 소비지출을 파악할 수 있는 분기별 데이터를 수집한다. 소비패턴은 이러한 데이터로부터 아래 <표 1>과 같은 변수를 이용하여 정의될 수 있다.

<표 1> 소비패턴 관련 변수

유형	변수
고객	성별, 연령대, 지역 등
결제	날짜, 품목, 금액, 빈도 등
가맹점	위치, 지역 특성, 상권의 특성, 가맹점의 특성 등

2. 전체가계의 소비패턴

전체가계를 대상으로 이루어진 선행연구들은 가계를 분류하지 않고 소비항목만 분류한 후 항목별 소비총액을 분석하였다. McCully, C. P. (2011)는 미국 경제분석국의 개인소비지출 데이터를 이용해 시대에 따른 소비총액의 증감을 분석하였다. 22개의 소비항목의 1959년부터 2009년까지 장기간에 걸친 소비비중 증감량을 구하고 그 원인을 제시하였다. 최지혜 외 (2021)는 국내 서울시 신용카드 결제 데이터

를 가지고 코로나19 기간동안의 소비패턴 변화를 분석하였다. 63개 소비항목을 기준으로 2018년부터 2020년까지의 오프라인 가맹점에서 발생한 매출액 변화를 통해 업종별로 코로나19의 영향을 유형화하였다. 이러한 연구들은 모두 사회경제적 변화에 따른 소비행동을 설명할 수 있지만, 가계의 다양성은 반영하지 못하였다.

3. 가계유형별 소비패턴

가계의 다양성을 반영한 소비패턴을 연구하는 방식은 가계를 소득 등 인구통계학적 특성에 따라 분류하고 각 가계의 소비행동을 파악하는 것으로, 소비패턴에 관한 연구의 대부분이 여기에 해당한다. Baker, S. 외 (2020)는 가계를 인구통계학 변수로 분류한 후 코로나19 발생 전후의 소비항목별 지출금액 증감을 통해 가계의 다양성을 설명하였다. 이와 달리 본 연구에서는 소비항목별 소비비중을 통해 가계의 다양성을 정의하였다. Chai, A. 외 (2015)는 가계 소비의 다양성을 측정하였고, 가계의 소득이 늘어날 수록 다양성이 증가하는 것을 발견하였다. 해당 연구에서는 같은 인구통계학적 특징을 가진 두 가계가 다른 소비패턴을 보이는 것을 다양성으로 정의하였다. 본 연구에서는 인구통계학적 특징과 상관 없이 가계의 소비비중이 다를 경우 다양성을 가지는 것으로 보았다. Schanzenbach, D. 외 (2016)는 가계 소비를 30년 전후로 비교하여 소비패턴의 변화를 연구하였다. 해당 연구에서는 가계를 소득에 따라 분류한 후 주요 소비항목 별 지출과 예산을 비교하였고, 그 결과 특정 소비항목에서 소득 별 차이가 발생함을 밝혀 냈다. 이 외에도 소비자의 성별, 지역, 소득과 같은 인구통계학 변수에 따른 주요 소비항목별 소비 변화를 분석하거나 (Chronopoulos, D. 외, 2020), 가계를 인구조사 지역별로 구분한 다음 소비항목별 소비 비중를 살펴 봄으로써 소비패턴을 정의한 연구가 있다(Yocum, C., 2007). 이러한 연구들은 모두 소비자를 인구통계학적 특성으로 유형화하여 소비패턴을 선형적으로 분석한 것으로, 다양한 가계의 복잡하고 비선형적인 소비패턴을 잡아내지 못 하였다는 한계가 있다.

또한 이상의 연구들보다 소비자나 소비항목을 더욱 세분화하여 특정 소비자 그룹 혹은 특정 소비항목의 소비패턴을 분석한 연구들이 존재한다. 최옥금 (2011)은 노인가구를 대상으로 소비지출을 유형화하였고, 정원오 외 (2011)는 빈곤계층을 대상으로 소비패턴을 연구하였다. Chen, C. 외 (2014)는 미국 주별 레지던트 연수지역을 소비수준에 따라 나눈 다음, 지역마다 의료비 지출에 차이가 발생함을 보여주었다. Douglas, N. 외 (2004)는 태평양 섬 항구의 유람선 승객들을 성별, 나이, 소득 등으로 분류한 다음 유람선 관련 소비를 분석하였다. Davis, M. 외 (2016)는 노인계층의 의료비 관련 소비패턴을 말년의 소비량 추이를 기준으로 나누어 네 가지 유형으로 제시하였다.

4. 소비행동별 소비패턴

소비지출에 따라 가구를 유형화하는 연구들은 주로 군집분석을 이용하여 가구를

몇 개의 군집으로 분류한 후 이들의 소비특성을 살펴보거나 각 군집에 영향을 미치는 요인을 찾아낸다. 초기의 연구인 손상희(1993)는 미국 노동통계국의 소비지출조사 자료를 이용하여 미국가구를 유형화하였는데, 총 31개의 소비변수를 계층적 군집화 알고리즘으로 군집분석 하였다. 최근의 연구인 성영애(2013)는 국내 1인 가구의 연령대별 소비지출패턴을 제시하였다. 군집분석에는 총 12개의 소비변수를 사용하였고 k-평균 군집화 알고리즘을 활용하였다. 이상의 연구들은 가계를 소비행동을 기준으로 군집화하였지만 고차원의 데이터에서 잘 작동하기 어려운 전통 군집화 알고리즘을 사용하였다는 한계가 존재한다.

5. 기계학습 알고리즘

1) 기계학습

기계학습 알고리즘은 데이터로부터 학습함으로써 어떤 작업에 대한 성능이 향상되는 알고리즘을 말한다(Tom Mitchell, 1998). 기계학습은 목적에 따라 지도학습(Supervised Learning)과 비지도학습(Unsupervised Learning)으로 나뉜다. 지도학습은 입력과 출력 데이터쌍이 주어졌을 때 둘의 관계를 설명하는 함수를 근사하는 것으로, 주로 예측 작업을 위해 사용된다. 비지도학습은 입력 데이터만 주어졌을 때 그것 자체의 관계를 통해 데이터의 구조를 파악하는 것으로, 주로 데이터에 담긴 정보를 압축하거나 잘 설명하기 위해 사용된다.

2) 군집화

군집화는 대표적인 비지도학습 기법 중 하나로, 데이터 포인트들을 '군집'이라고 불리는 유사성 그룹으로 묶는 것을 말한다(David, S., 2013). 군집화 알고리즘은 같은 군집으로 묶인 데이터들은 특성이 비슷하고 다른 군집으로 묶인 데이터들은 특성이 서로 다른 것을 목적으로 학습한다. 군집화는 데이터에서 패턴을 찾아내는 데 유용하기 때문에 고객 세분화, 이메일 그룹화, 이미지 분할 등의 작업에서 사용된다. 군집을 만들기 위해서는 데이터 포인트 사이의 유사도를 측정해야 하는데, 정량적인 데이터 특성을 비교하기 위해서는 거리를 주로 사용한다. 본 연구에서는 가장 널리 쓰이는 거리함수인 민코프스키 거리(Minkowski distance) 중 유클리드 거리(Euclidean distance)를 사용하였다. 아래 <표 2>는 유클리드 거리를 비롯한 민코프스키 거리의 여러 정의를 나타낸다.

<표 2> 민코프스키 거리

거리함수	수식	거리에 대한 정의
민코프스키 거리	$(\sum_{l=1}^{d}\|x_{il}-x_{jl}\|^{n})^{l/n}$	$(n=1)$ 맨해튼 거리 $(n=2)$ 유클리드 거리 $(n=3)$ 체비쇼프 거리

전통적인 군집화 알고리즘은 군집화 방식에 따라 크게 분할, 계층, 분포 등의 범주로 나눌 수 있다(Xu, D. 외, 2015). 예를 들어 가장 대표적인 분할 기반 알고리즘인 k-평균 군집화(k-means clustering)를 비롯하여 계층 기반 알고리즘인 병합 군집화(Agglomerative Clustering), 밀도 기반 알고리즘인 DBSCAN, 그리고 분포 기반 알고리즘인 GMM 등이 있다. 이러한 방법들은 모델 구조가 간단하다는 장점이 있지만, 고차원 데이터에서는 잘 작동하지 않는다는 문제가 있다.

최근에는 고차원 데이터를 효과적으로 다루기 위해 심층 학습(Deep Learning)을 이용한 심층 군집화(Deep Clustering)가 활발히 연구되고 있는데, 주로 표현 학습(Representation Learning)과 군집화를 결합한 형태의 알고리즘이다. 즉, 고차원 데이터의 대표 특징을 추출한 후 추출된 저차원의 특징을 가지고 군집화하는 것이다. 예를 들어 합성곱 신경망(Convolutional Neural Networks, CNN)이나 오토인코더(Autoencoder)로 대표 특징을 추출한 후 여러 알고리즘으로 군집화 하거나(Yang, J. 외, 2016; Xie, J. 외, 2016), 표현학습과 k-means, VAE, HGMM 군집화를 동시에 최적화하는 DTCR(Ma, Q. 외, 2019), VaDE(Jiang, Z. 외, 2016), HCRL(Shin, S. 외, 2020) 알고리즘이 있다. 이러한 방법들은 고차원 데이터에서 잘 작동하지만, 모델 구조가 복잡해서 구현과 설명이 어렵다는 단점이 있다.

3) 심층학습

심층학습 또는 딥러닝으로 불리는 신경망 알고리즘은 서로 연결된 뉴런 계층 구조와 역전파를 이용해 원하는 함수를 찾는 학습 방식으로, 전통적인 기계학습에 비해 두 가지 큰 차이점을 갖는다. 첫째, 심층학습은 비선형 관계를 파악할 때 유리하다. 신경망 모델은 크게 입력층, 은닉층, 출력층으로 이루어진 구조인데, 다른 층으로 연결된 변수들을 활성함수를 이용해 비선형적인 방법으로 결합해서 잠재변수를 만든다. 그리고 이러한 과정이 우리가 근사하는 함수를 비선형으로 만들고, 비선형성이 더해진 함수는 더욱 복잡한 데이터 관계를 포착할 수 있게 된다. 둘째, 심층학습은 계산 비용이 비싸다. 기계학습 알고리즘에 비해 훨신 많은 파라미터를 학습시켜야 하는 딥러닝 알고리즘은 높은 성능의 컴퓨팅 파워와 많은 양의 데이터를 요구한다. 따라서 현실에서 활용하기 위해서는 모델 사이즈를 고려해야 하고, 간단한 구조로 좋은 성능을 얻는 모델을 만드는 것이 중요하다.

Ⅲ. 연구방법

1. 연구문제

본 연구의 목적은 군집화를 이용하여 우리나라 가구 소비활동의 다양성을 충분히 반영하는 소비패턴을 찾고 분석하는 것이다. 따라서 소비패턴은 각 소비항목이 전체소비에서 차지하는 비중을 기준변수로 하여 가구들을 군집화한 결과로 형성된 군집들을 말한다. 소비금액이 아닌 소비비중을 기준변수로 사용하는 이유는 가구마다

지출규모가 다를 경우 소비패턴이 아닌 지출금액에 따라 군집화될 수 있기 때문이다. 즉, 변수의 저마다 다른 측정단위 크기에 상관없이 그 차이에 따라 일정하게 거리를 측정할 수 있도록 한다. 이렇게 소비패턴이 소비항목별 소비비중에 따라 유형화될 것이라는 가정 하에 다음과 같이 연구문제를 설정하고 소비패턴을 탐색하였다.

1) 국내 가구의 소비패턴은 어떻게 분류되는가?

각 군집 별 가구 실태 관련 변수와 소비지출 관련 변수의 기술통계량을 통해 소비패턴을 비교분석한다. 소비특성 따라 나뉜 군집들은 소비지출 항목들에 대해 서로 구별되는 분포를 가질 것이다.

2) 각 소비패턴에 영향을 주는 요인은 무엇인가?

로지스틱 회귀모형을 이용해 각 군집 별 가구의 사회인구학적 특성을 파악한다. 가계동향조사 데이터에는 소득, 가구주 연령, 가구원수 등 다양한 가구 실태 관련 변수가 존재하는데, 이러한 변수들을 독립변수로 사용하고 가구가 어떤 군집에 속할 여부를 종속변수로 사용하면 각 군집이 어떤 사회인구학적 변수와 밀접한 관련이 있는지 알 수 있다.

3) 각 소비패턴은 가구의 다양성을 잘 반영하였는가?

다중 곡선 차트를 이용해 가계지출금액에 따른 소비 배분을 군집 별로 파악한다. 만약 군집들이 가구의 다양성을 포착하였다면, 군집마다 소비 배분의 모양과 추세가 서로 잘 구분될 것이다.

2. 연구자료

본 연구는 우리나라 가계의 소비패턴을 파악하기 위해 통계청 가계동향조사 자료 중 소비지출 항목을 사용하였다. 가계동향조사는 대한민국에 거주하는 모든 일반가구 중에서 매월 7,200여 가구를 대상으로 가계수지 실태를 파악한다. 이 중 가계지출은 한 가구를 형성하는 가족이 경제활동을 하는 데 지출하는 금액의 총합으로, 소비지출과 비소비지출로 나뉜다. 소비지출은 식료품비나 주거비 등 일상생활에 필요한 상품이나 서비스를 구입한 비용이다. 비소비지출은 각종 세금과 이자비용 등 소비와 직접관련이 없는 지출을 말한다.

본 연구에서는 시간에 따라 변화하는 소비 성향을 소득과 연관지어 분석하기 위해, 새 전용표본 가구를 대상으로 소득과 지출의 통합조사가 시작된 2019년부터 최근까지 3년 동안의 데이터(2019~2021년)를 사용하였다. 가계동향조사 자료에 존재하는 여러 변수는 용도에 따라 아래 <표 3>과 같이 구분되었다. 먼저 각 가구의 소비 특성을 기준으로 가구들을 군집화하기 위해, 12개 소비지출 항목들을 군집화에 사용하였다. 그리고 형성된 군집들에 속한 가구들의 특성을 파악하기 위해, 가구 실태와 가계소득 관련 항목들을 사후분석에 활용하였다.

<표 3> 변수의 구분

용도	변수
군집화	식료품 및 비주류음료, 주류 및 담배, 의류 및 신발, 주거 및 수도광열, 가정용품 및 가사서비스, 보건, 교통, 통신, 오락 및 문화, 교육, 음식 및 숙박, 기타 상품 및 서비스
사후분석	소득, 가구주 성별, 가구주 연령, 가구주 학력, 자동차보유대수, 주택소유유무, 가구구분, 가구원수, 취업인원수, 세대구분

3. 방법론

1) 군집화

본 연구에서 사용된 군집화 모형은 McConville 외 (2021)가 제안한 Not Too Deep Clustering(N2D)이다. N2D는 심층 표현학습 기법과 전통 군집화 기법을 결합한 모형으로, 오토인코더(Autoencoder)와 매니폴드(Manifold) 학습 기법을 통해 데이터의 비선형 관계를 파악하여 차원을 축소한 후 군집화를 수행한다. 이는 본 연구에서 사용하는 고차원 데이터를 효과적으로 다루기에 적합하며, 기존 심층 군집화 알고리즘보다 모델구조가 간단하여 학습 속도가 빠르면서도 성능이 좋은 장점이 있다. 즉, 전통 군집화 알고리즘과 심층 군집화의 장점을 합친 모델이라고 할 수 있다.

(1) 오토인코더

N2D모형은 먼저 오토인코더 알고리즘을 이용해 데이터의 군집 개수 만큼의 차원까지 축소한다. 오토인코더는 인코더(Encoder)와 디코더(Decoder)로 구성된 심층 신경망이다. 인코더는 입력 데이터를 압축된 특성으로 매핑하고, 디코더는 압축된 특성을 다시 원래의 입력 공간으로 매핑한다. 그리고 인코더와 디코더는 원본 입력 데이터와 생성된 입력 데이터 사이의 손실을 최소화하는 방향으로 학습된다. 즉, 오토인코더의 목적은 학습이 완료된 인코더를 이용해 입력 데이터를 압축된 특성으로 변환하는 것이다.

(2) 매니폴드 학습

N2D모형은 오토인코더를 통해 축소된 입력 데이터를 매니폴드 기법을 이용해 3차원까지 한 번 더 축소한다. 매니폴드란 고차원 데이터의 정보를 잘 표현할 수 있는 저차원의 데이터 공간을 말한다. 본 연구에서는 UMAP(Uniform Manifold Approximation and Projection) 알고리즘을 이용해 매니폴드를 학습하였다. UMAP은 고차원 데이터의 지역적 구조와 전역적 구조를 모두 잘 유지한다는 특징이 있기 때문에 UMAP을 통해 만들어진 압축된 데이터는 군집성이 우수하다.

(3) 계층적 군집화 및 병합 군집화

N2D모형 내에서 사용한 전통 군집화 기법은 계층적 군집화(Hierarchical Clustering)이다. 계층적 군집화는 데이터로부터 계층적 관계를 파악함으로써 계층적 구조를 가진 군집들을 만들어 내는 알고리즘이다. 군집은 모든 데이터포인트 쌍 사이의 거리를 계산하는 탐욕적 방식(Greedy manner)으로 계산된다. 본 연구에서 사용된 계층적 군집화 알고리즘인 병합 군집화(Agglomerative Clustering)는 각 데이터포인트가 하나의 독립된 클러스터에서 출발하여 단 하나의 클러스터가 될 때까지 비슷한 데이터포인트 혹은 군집을 병합하는 알고리즘이다. 본 연구에서는 병합되는 군집의 분산을 최소화하는 와드 연결법(Ward linkage)을 거리 측정 방식으로 사용하였다.

2) 군집화 성능평가

데이터에 군집 정보에 대한 정답이 존재할 경우, 모델 성능은 평가용 데이터셋의 예측 군집과 실제 군집을 비교하는 혼동행렬(Confusion matrix), 랜드지수(Rand Index, RI) 등의 지표를 이용한 외적 평가가 가능하다. 즉, 형성된 군집이 얼마나 정답에 가까운지를 평가하는 것이다. 그러나 본 연구에서 사용한 데이터에는 실제로 정해진 군집이 없으며 새롭게 소비패턴을 정의하는 문제를 풀고자 하기 때문에 내적 평가를 하였다. 내적 평가는 군집화에 사용된 데이터의 유사도를 비교하는 실루엣 계수(Silhouette coefficient), DB지수(Davies-Bouldin Index, DBI) 등의 지표를 이용하여 알고리즘의 타당성을 평가하는 것이다. 군집의 응집도와 분리도를 측정하는 실루엣 계수는 -1 에서 1 사이의 값을 가지며, 값이 클 수록 군집화가 잘 이루어졌다고 말할 수 있다. 군집의 평균 유사도를 측정하는 DB지수는 0 이상의 값을 가지며, 값이 작을 수록 군집화가 잘 이루어졌다고 말할 수 있다.

3) 로지스틱 회귀모형

본 연구는 각 소비패턴의 가구 특성을 설명하기 위해 로지스틱 회귀모형(Logistic Regression)을 사용하였다. 로지스틱 회귀모형은 로짓(Logit)함수를 이용해 특정 확률을 모델링하는 분류모델로, 설명변수에 대한 반응변수를 설명하기 위해 사용된다.

Ⅳ. 연구결과

1. 군집화 성능

본 연구에서 사용된 계층적 군집화 알고리즘은 병합 군집화로, 군집 개수를 사용자가 직접 지정해 주어야 한다. 하지만 소비패턴을 찾는 것은 비지도학습 문제로, 군집 개수에 대한 사전 정보가 없이 적절한 군집 개수를 찾아야 할 필요가 있다. 최적의 군집 개수를 결정하는 한 가지 방법은 군집 개수별 평가지표를 비교한 후 그 중 가장 우수한 성능을 보이는 경우를 선택하는 것이다(Rousseeuw, P., 1987). 이에 본 연구는 아래 <표 4>와 같이 6~12개의 군집 개수 별 평가지표를 비교한 후 가장 좋은 성능을 보인 군집 개수 9개를 선택하였다.

<표 4> 군집 개수에 따른 군집화 성능

군집 개수	평가지표	
	실루엣 계수(↑)	DB지수(↓)
6	0.062	2.375
7	0.052	2.517
8	0.113	1.948
9	**0.138**	**1.888**
10	0.089	2.159
11	0.085	2.154
12	0.023	3.065

실루엣 계수와 DB지수와 같은 내적 평가지표는 유사도를 기반으로 모델에 데이터가 얼마나 잘 적합되었는지를 보기 위한 참고 지표이며, 알고리즘의 성능을 절대적으로 평가할 수는 없다(Estivill, V., 2002). 따라서 사후 분석으로 기술통계량, 로지스틱 회귀모형, 소비 포트폴리오 차트 등을 이용하여 군집을 비교분석하였다.

2. 소비패턴

<표 5>는 군집결과 나타난 9개 군집의 가구실태와 소비지출 특성을 보여준다. 분석의 편의를 위해 군집 번호(1~9)는 평균 소득이 높은 순으로 부여되었다. 패널 A에는 군집 별로 10가지의 가구실태 관련 변수의 평균이 표기되었고, 패널 B에는 군집 별로 12가지 소비항목에 대한 소비비중의 평균과 표준편차가 표기되었다. 중요 발견점을 표시하기 위해, <표 5>의 모든 값들은 항목을 기준으로 상위 2위의 값까지 굵게 표시되었다. 즉, 같은 항목에 대해 어떤 군집이 다른 군집들보다 높은 값을 가지면 굵게 표시하였다.

<표 5> 군집 별 가구실태 및 소비비중

패널A. 가구실태

군집	월소득(원)	가구주성별	가구주연령(세)	가구주학력	자동차보유대수(대)
1	**7,050,470**	1.13	51	**4.78**	**1.43**
2	**6,151,896**	1.19	46	**5.30**	**1.19**
3	5,182,612	1.28	52	4.67	1.06
4	4,684,319	1.24	50	4.66	1.00
5	3,833,847	1.32	57	4.20	0.83
6	3,152,114	1.30	45	4.72	0.59
7	2,980,320	1.35	**66**	3.50	0.58
8	2,355,382	**1.42**	**69**	3.22	0.40
9	1,947,374	**1.49**	63	3.55	0.27
		남성1, 여성2		*무학1 ~ 박사8*	

군집	주택소유유무	가구구분	가구원수(명)	취업인원수(명)	세대구분(세대)
1	1.24	1.34	**2.64**	**1.55**	**1.60**
2	1.26	1.29	**3.53**	**1.49**	**1.98**
3	1.28	1.44	2.28	1.36	1.47
4	1.34	1.36	2.23	1.47	1.48
5	1.33	1.49	2.24	1.20	1.45
6	**1.81**	1.33	1.53	1.07	1.21
7	1.22	1.64	1.94	0.97	1.26
8	1.17	**1.69**	1.85	0.85	1.22
9	**1.71**	**1.67**	1.47	0.62	1.16
	있음1, 없음2	*근로자가구1, 근로자외가구2*			

패널B. 소비비중

군집	식료품	주거	음식숙박	보건	교통	기타
1	0.031 (0.019)	0.020 (0.020)	0.030 (0.020)	0.016 (0.021)	**0.825** (0.066)	0.021 (0.015)
2	0.149 (0.079)	0.093 (0.058)	0.125 (0.057)	0.057 (0.054)	0.070 (0.145)	0.073 (0.143)
3	0.141 (0.065)	0.085 (0.065)	0.102 (0.060)	0.057 (0.059)	**0.136** (0.041)	**0.160** (0.038)
4	0.163 (0.077)	0.097 (0.053)	**0.234** (0.095)	0.049 (0.043)	0.122 (0.071)	**0.097** (0.056)
5	**0.202** (0.078)	0.146 (0.087)	0.110 (0.062)	**0.119** (0.084)	0.077 (0.048)	0.083 (0.049)
6	0.105 (0.087)	**0.286** (0.065)	**0.224** (0.050)	0.033 (0.147)	0.093 (0.041)	0.057 (0.042)
7	0.179 (0.053)	0.096 (0.071)	0.063 (0.103)	**0.417** (0.037)	0.050 (0.058)	0.055 (0.042)
8	**0.422** (0.104)	0.124 (0.076)	0.069 (0.055)	0.101 (0.079)	0.047 (0.040)	0.065 (0.045)
9	0.187 (0.094)	**0.476** (0.114)	0.062 (0.055)	0.065 (0.070)	0.037 (0.035)	0.038 (0.032)

군집	통신	오락문화	의류신발	가정용품	교육	주류담배
1	0.012 (0.012)	0.012 (0.016)	0.010 (0.011)	0.007 (0.010)	**0.013** (0.026)	0.004 (0.006)
2	0.054 (0.065)	**0.058** (0.139)	0.054 (0.071)	0.039 (0.053)	**0.217** (0.033)	0.012 (0.030)
3	0.056 (0.046)	**0.133** (0.048)	**0.061** (0.046)	0.040 (0.053)	**0.013** (0.113)	0.016 (0.018)
4	**0.064** (0.053)	0.050 (0.045)	**0.058** (0.059)	0.034 (0.044)	0.006 (0.019)	**0.026** (0.040)
5	**0.057** (0.051)	0.044 (0.044)	0.049 (0.053)	**0.087** (0.135)	0.006 (0.018)	0.021 (0.039)
6	**0.057** (0.033)	0.044 (0.038)	0.039 (0.037)	0.023 (0.053)	0.003 (0.015)	**0.034** (0.025)
7	0.033 (0.051)	0.028 (0.048)	0.027 (0.048)	0.035 (0.034)	0.003 (0.015)	0.012 (0.047)
8	0.042 (0.038)	0.034 (0.039)	0.034 (0.044)	**0.044** (0.056)	0.002 (0.010)	0.017 (0.036)
9	0.037 (0.033)	0.025 (0.029)	0.018 (0.028)	0.032 (0.054)	0.003 (0.015)	0.021 (0.044)

*값: 평균(표준편차)

위 <표 5>를 기준으로 요약한 군집별 소비특성은 다음과 같다. 단, 평균값이 높더라도 표준편차가 지나치게 높은 경우 평균값이 그 특성을 대표한다고 보기 힘들기 때문에 소비특성에서 제외하였다.

1) 군집 1: 교통비 중심형

교통비 지출비중이 82.5%로 과도하게 높아 다른 소비항목에는 거의 지출하지 않는다. 이 군집의 월소득과 자동차보유대수는 전체 군집 중 가장 높기 때문에 높은 소득을 바탕으로 자동차를 구매하는 소비패턴이 포착된 것으로 보인다. 남성의 비중이 여성보다 더 높고, 가구주 연령은 51세이며 가구주 학력은 대학교(3년제 이하)로 전체 군집 중 두 번째로 높다.

2) 군집 2: 교육비 중심형

교육비 지출비중이 21.7%로 다른 군집에 비해 월등히 높고 그 다음으로는 오락문화비, 의류신발비 등의 비중이 높다. 이 군집은 월소득이 두 번째로 높고, 특히 가구주 학력이 대학교(4년제 이상)로 가장 높다. 또한 가구주 연령이 낮은 편이면서 세대 수와 가구원수가 가장 많기 때문에 학교와 학원에서 공부하는 자녀들에게 사용하는 지출이 반영된 집단이라고 할 수 있다.

3) 군집 3: 여가생활(문화) 중심형

오락문화비, 의류신발비, 교통비, 기타 지출비중이 다른 군집에 비해 높다. 이 군집은 소득, 가구주 학력, 자동차보유대수가 모두 높은 편으로 안정된 소득을 바탕으로 구매 중심의 다양한 여가생활을 즐기는 가계의 소비패턴으로 볼 수 있다. 나머지 소비항목에 대해서는 골고루 소비비중이 배분되어 있는 편이다.

4) 군집 4: 여가생활(여행) 중심형

음식숙박비, 통신비, 기타 지출비중이 다른 군집에 비해 높다. 특히, 음식숙박비 지출비중이 23.4%로 다른 군집에 비해 월등히 높은데 이는 외식과 여행을 많이 하였음을 보여준다. 또한 통신비와 기타 지출비중이 각각 6.4%, 9.7%로 비교적 매우 높은 것이 또다른 특징이다. 이 군집의 가구주 연령은 두 번째로 낮다.

5) 군집 5: 가정 중심형

식료품비, 보건비, 통신비, 가정용품 지출비중이 다른 군집에 비해 높은데, 이러한 소비항목들은 모두 가정 내 가구원들을 위해 주로 쓰인다는 공통점이 있다. 이 군집의 가구주 연령, 가구주 학력, 자동차 보유대수, 주택소유유무 등은 모두 중간 정도이며 가구원수는 많은 편이다.

6) 군집 6: 복합형

이 군집은 다른 군집들과 달리 다양한 소비항목에 골고루 소비하는 유형이다. 주거비, 음식숙박비, 주류담배비, 통신비 지출비중이 다른 군집에 비해 높다. 대신에

보건비와 교육비는 가장 낮은 수준이다. 이 유형의 특징은 가구주 연령이 가장 낮으며 주택을 소유하지 않는 비중이 높다.

7) 군집 7,8,9: 보건 중심형, 식료품 중심형, 주거 중심형

세 군집은 모두 한 소비항목에 편중되어 소비하는 유형으로, 각각 보건비(41.7%), 식료품비(42.2%), 주거비(47.6%)에 지출하는 비중이 과도하게 높다. 이 군집들은 가구주 연령이 제일 높고 월소득, 가구주 학력, 자동차 보유대수, 취업인원수가 가장 낮다는 특징이 있다. 또한 근로자외가구 비중이 가장 높다. 따라서 이 소비패턴은 낮은 소득으로 인해 생존에 가장 필수적인 소비항목에 대부분을 소비한 뒤 다른 소비는 하지 못하는 집단을 나타낸다.

3. 소비패턴에 영향을 미치는 요인

<표 6>은 군집을 종속변수로 설정하고 월소득, 가구주 성별, 가구주 연령, 가구주 학력, 자동차 보유대수, 주택소유유무, 가구원수를 독립변수로 하여 로지스틱 회귀분석한 결과를 요약한 것이다. 각 변수에 대한 회귀계수(Coefficient)와 통계적 유의성이 오즈비(Odds ratio)와 함께 표기되어 있다. 각 군집에 해당하는 중요 변수들은 양의 관계일 경우 파란색, 음의 관계일 경우 빨간색으로 표시되었다. 분석결과는 가계 다양성과 가구실태 변수 사이에 강한 관계가 있음을 보여준다.

<표 6> 군집별 로지스틱 회귀분석 결과

변수	군집1		군집2		군집3		군집4		군집5		군집6		군집7		군집8		군집9	
	Coeff	Odds ratio	Coeff	Odds ratio	Coeff	Odds ratio	Coeff	Odds ratio	Coeff	Odds ratio	Coeff	Odds ratio	Coeff	Odds ratio	Coeff	Odds ratio	Coeff	Odds ratio
Constant	-7.853***	0.000	-3.963***	0.019	-2.620***	0.073	-1.508***	0.221	-2.446***	0.087	-2.641***	0.071	-2.992***	0.050	-2.263***	0.104	-3.352***	0.035
월소득 (175만원 이하)																		
176~335만원	-0.185	0.831	0.132*	1.141	0.299***	1.349	0.613***	1.846	0.065*	1.067	0.415***	1.514	-0.076*	0.927	-0.344***	0.709	-0.582***	0.559
336~550만원	0.288	1.334	0.278***	1.320	0.370***	1.448	0.868***	2.382	0.010	1.010	0.115*	1.122	-0.188***	0.829	-0.574***	0.563	-0.922***	0.398
551만원 이상	0.848***	2.335	0.480***	1.616	0.573***	1.774	0.841***	2.319	-0.242***	0.785	-0.093	0.911	-0.291***	0.748	-1.067***	0.344	-0.841***	0.431
가구주 성별 (여성)	-0.264	0.768	0.345***	1.412	0.185***	1.203	-0.216***	0.806	0.217***	1.242	-0.635***	0.530	0.058	1.060	0.244***	1.276	-0.041	0.960
가구주 연령 (43세 이하)																		
44~55세	-0.112	0.894	0.037	1.038	-0.225***	0.799	-0.134***	0.875	0.169***	1.184	0.029	1.029	0.349***	1.418	0.457***	1.579	0.403***	1.496
56~65세	-0.323*	0.724	-1.684***	0.186	-0.020	0.980	-0.087**	0.917	0.654***	1.923	-0.373***	0.689	0.924***	2.519	1.401***	4.059	0.834***	2.303
66세 이상	-0.437*	0.646	-2.239***	0.107	-0.556***	0.573	-1.092***	0.336	0.354***	1.425	-1.198***	0.302	1.444***	4.238	2.035***	7.652	1.290***	3.633
가구주 학력 (초등학교 이하)																		
중학교	-0.057	0.945	0.463***	1.589	0.055	1.057	0.160***	1.174	0.074*	1.077	0.408***	1.504	-0.084*	0.919	-0.113***	0.893	0.096*	1.101
고등학교	-0.362*	0.696	0.881**	2.413	0.071	1.074	0.034	1.035	-0.025	0.975	0.299***	1.349	-0.276***	0.759	-0.166***	0.847	0.092	1.096
대학교 이상	-0.361	0.697	1.217***	3.377	0.245***	1.278	-0.263***	0.769	-0.111	0.895	0.155	1.168	-0.239*	0.787	-0.119	0.888	0.016	1.016
자동차 보유대수 (0대)																		
1대	3.470***	32.137	0.119**	1.126	0.608***	1.837	0.428***	1.534	0.225***	1.252	-0.039	0.962	-0.214***	0.807	-0.758***	0.469	-0.765***	0.465
2대 이상	4.084***	59.383	0.032	1.033	0.942***	2.565	0.547***	1.728	0.092*	1.096	-0.226**	0.798	-0.453***	0.636	-1.367***	0.255	-0.929***	0.395
주택소유유무 (없음)	-0.085	0.919	-0.303***	0.739	-0.333***	0.717	-0.113***	0.893	0.136***	1.146	1.671***	5.317	-0.449***	0.638	-1.052***	0.349	1.646***	5.186
가구원수 (1명)																		
2명	-0.345*	0.708	0.882***	2.416	-0.214***	0.807	-0.354***	0.702	0.504***	1.655	-0.838***	0.433	0.387***	1.473	0.489***	1.631	-0.248***	0.780
3명 이상	0.699***	2.012	2.650	14.154	-0.641***	0.527	-0.850***	0.427	0.396***	1.486	-1.768***	0.171	-0.055	0.946	0.343***	1.409	-0.824***	0.439
가구수	447		11,333		6,886		15,632		12,170		5,602		5,619		10,167		4,374	

* < .05, ** < .01, *** < .001

<표 6>으로부터 도출된 군집별 주요 분석결과는 다음과 같다.

군집 1은 교통비 지출이 과도하게 많은 집단으로, 다른 군집들에 비해서 소득과 자동차 보유대수가 많은 가구들로 구성되어 있음을 알 수 있다. 또한 가구원수가 3명 이상으로 많은 가구일 가능성이 크다. 군집 2와 3은 소득이 많고 가구주가 비교적 젊으며 학력이 높고 주택을 소유한 가구들로 특징지을 수 있다. 가구원수는 1~2명 정도로 적을 가능성이 크다. 군집 4는 소득이 많은 편이며 가구주가 비교적 젊고 자동차와 주택을 보유한 가구로 구성되었다. 가구원수는 적을 가능성이 크다.

군집 5는 가구주 연령이 높은 편이고 자동차와 주택을 소유할 가능성이 적으면서 가구원수는 많은 가구를 나타낸다. 군집 6은 가구주 연령이 낮은 편이고 가구주 학력이 고등학교 이하인 가구들로 볼 수 있다. 주택을 소유하고 있지 않으며 가구원수는 적은 집단을 나타낸다. 군집 7, 8, 9는 공통적인 특징을 가지고 있다. 소득이 낮고, 가구주 연령이 높은 편이며, 가구주 학력은 낮고 자동차를 보유하고 있지 않는 가구들을 대표한다. 단, 군집 9가 군집 7과 8보다 더 적은 가구원수를 가질 가능성이 크다.

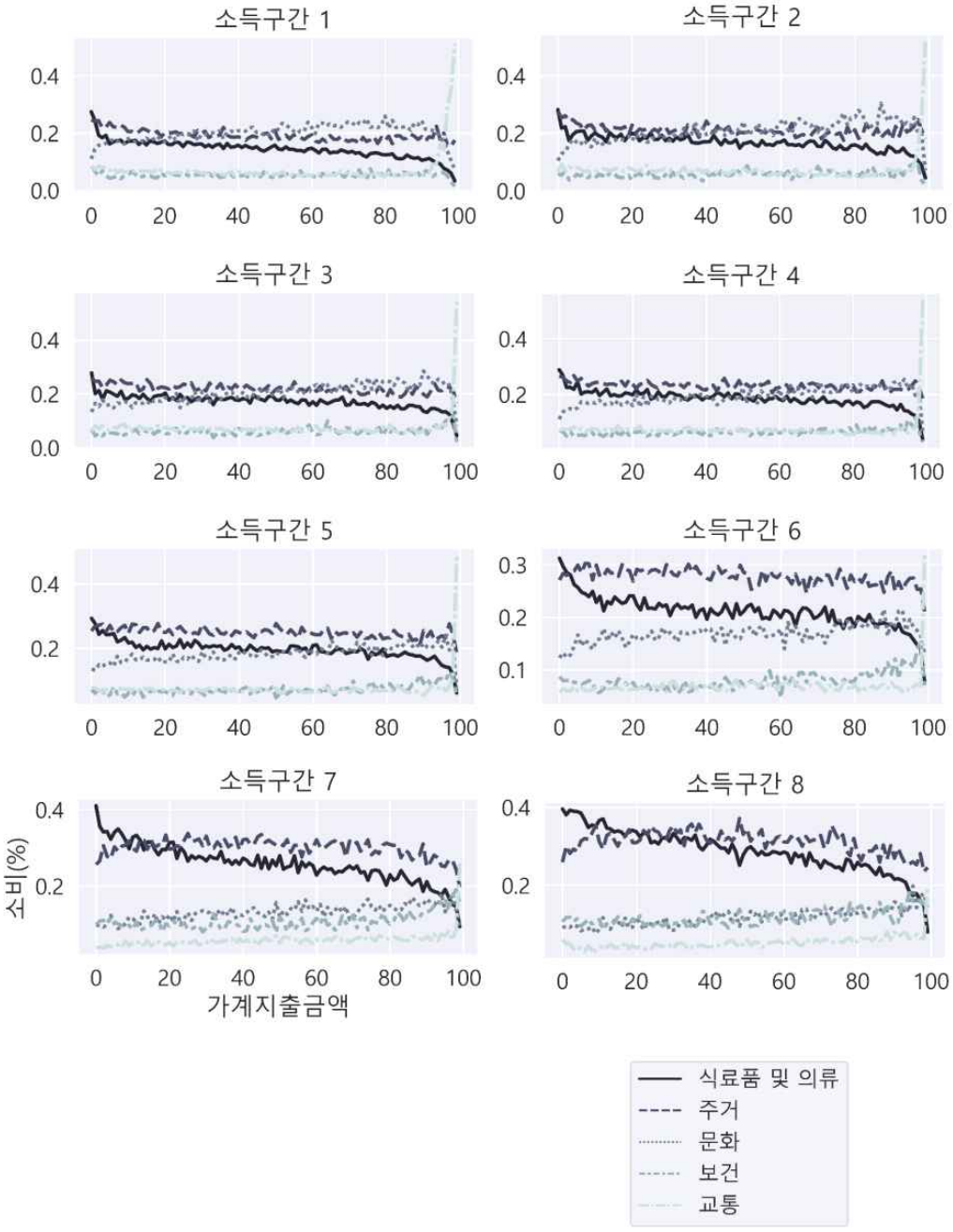
4. 소비패턴의 가구 다양성

아래 <그림 1, 2>는 가구 유형에 따른 소비 포트폴리오를 보여준다. 다중 곡선 차트에서 각 곡선은 가구들의 가계지출금액에 따른 한 소비항목에 대한 소비비중을 나타낸다. 즉, <그림 1, 2>는 같은 유형의 가구들을 가계지출금액에 따라 백분위로 나눈 다음 각 분위에 해당하는 가구들의 여러 소비항목에 대한 평균 소비비중을 계산한 값으로 차트를 그린 것이다. 이 때, 분석의 편의를 위해 소비항목은 군집화 기준변수인 12개를 비슷한 것끼리 병합하여 5개를 사용하였다. 소비항목 병합 기준은 다음과 같다.

- 식료품 및 의류: 식료품 및 비주류음료, 주류 및 담배, 의류 및 신발
- 주거: 주거 및 수도광열, 가정용품 및 가사서비스, 통신, 기타 상품 및 서비스
- 문화: 오락 및 문화, 음식 및 숙박, 교육
- 보건: 보건
- 교통: 교통

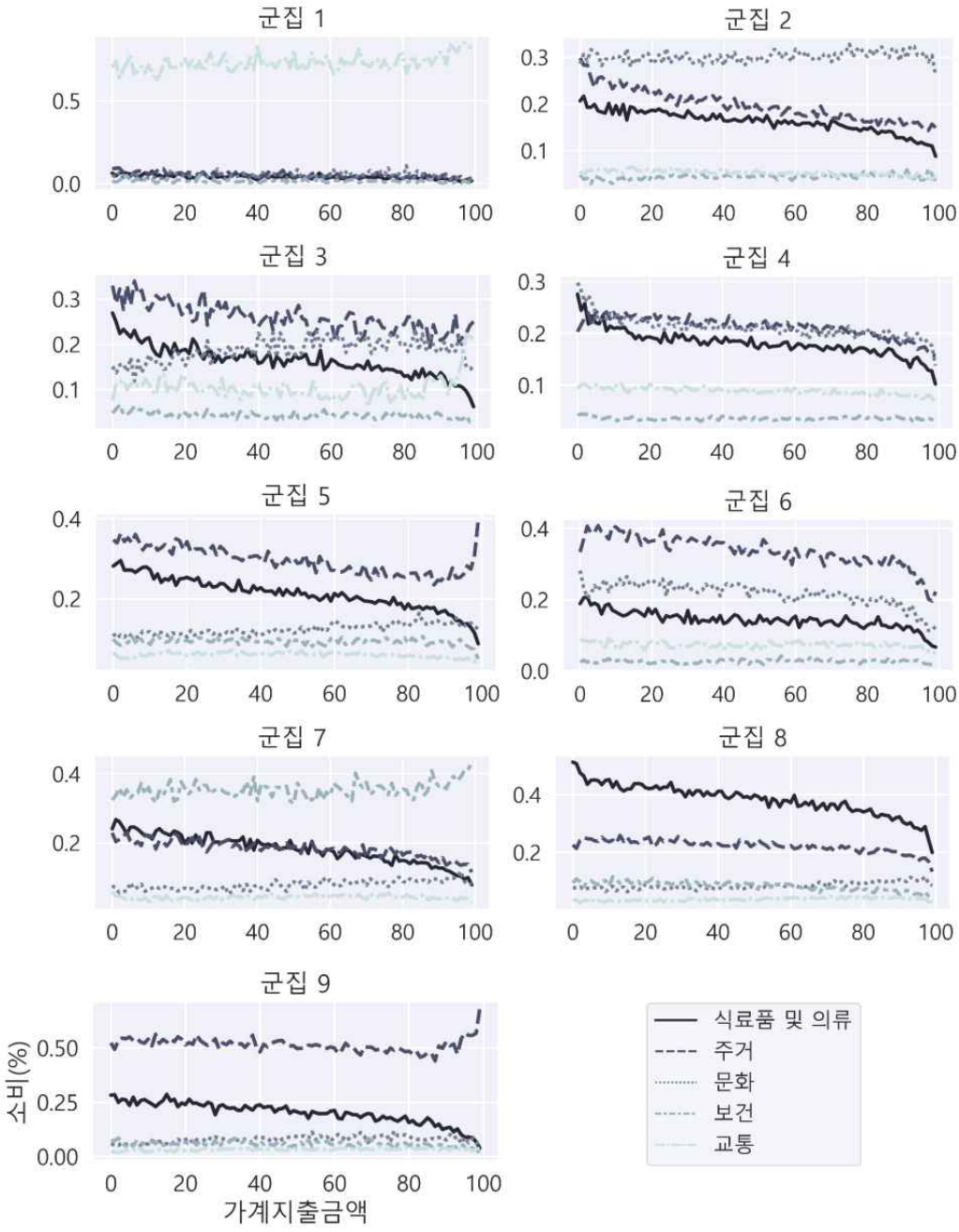
먼저 <그림 1>은 기존에 소비패턴을 분석할 때 주로 쓰이는 방식으로 소득을 기준으로 가구들을 유형화한 것이다. 전체 가구를 8개의 소득구간으로 분류한 후 소비 포트폴리오를 그렸을 때, 가구들의 소비 포트폴리오가 잘 구분되지 않는 것을 볼 수 있다. 예를 들어, 소득구간 1~5의 형태가 비슷하고 소득구간 7~8의 형태가 비슷하다.

<그림 1> 소득구간별 가구의 소비 포트폴리오



<그림2>는 본 연구에서 군집화한 방식으로 9개의 군집으로 가구들을 유형화한 것이다. 이 때, 각 군집에 속한 가구들의 소비 포트폴리오는 앞서 소득으로 가구들을 유형화했을 때와 달리 확연히 구분되는 것을 볼 수 있다. 예를 들어 군집 1은 교통비에 50%이상을 지출하고 나머지에는 비슷하게 지출하는 소비 포트폴리오를 보여 주고, 군집 2는 식료품 및 의류비와 주거비 소비비중이 가계지출금액이 커질수록 감소하는 소비 포트폴리오를 보여 준다. 이를 통해, 본 연구에서 활용한 심층 군집화 모형이 가계의 다양성을 잘 포착하였음을 확인할 수 있다.

<그림 2> 군집별 가구의 소비 포트폴리오



Ⅴ. 결론

가계의 소비활동은 가장 대표적인 금융 활동이다. 그리고 금융 활동은 개인화의 활용성과 가치가 크기 때문에 개인의 다양성을 고려하여 금융활동을 모델링하려는 연구가 이루어져왔다. 하지만 가계의 다양성은 흔히 인구통계학적 변수와 같은 변수로 구분되어 소비패턴과의 선형적 관계가 주로 분석되어 왔다. 소비지출 변수를 활용해 군집분석을 수행한 연구에서도 전통 기계학습 군집화 방법을 사용해 고차원 데이터의 복잡하고 비선형적인 관계를 포착하는 데 한계가 있었다. 최근에는 복잡한 데이터 관계를 포착하는 데 유리한 심층학습을 이용한 심층 군집화 방법이 활발히 연구되고 있는데, 그 중에서도 비교적 간단한 방법론을 사용하면 데이터에 과대적합될 위험을 피하면서도 모델 학습시간을 줄여 쉽게 활용할 수 있다.

따라서 본 연구는 간단한 모델구조를 가진 군집화 방법인 N2D 모형을 이용해 가계를 소비특성을 기준으로 군집화 함으로써 소비패턴을 찾고자 하였다. 군집화 평가지표인 실루엣 계수와 DB지수를 활용해 군집 개수를 결정하였고, 그 결과 다음과 같이 9개 군집으로 가구의 소비를 유형화하였다.

- 교통비 중심형, 교육비 중심형, 여가생활(문화) 중심형, 여가생활(여행) 중심형, 가정 중심형, 복합형, 보건 중심형, 식료품 중심형, 주거 중심형

또한 소비패턴에 영향을 미치는 요인을 파악하기 위해 가구실태변수를 활용한 로지스틱 회귀분석을 통해 다음과 같은 군집과의 관계를 파악하였다.

- 소득은 교통비 중심형, 교육비 중심형, 여가생활(문화) 중심형 소비패턴에서 높게 나타났다. 반면에, 한 소비항목에만 집중적으로 소비한 보건 중심형, 식료품 중심형, 주거 중심형 소비패턴에서는 소득이 낮았다.
- 가구주 성별은 교육비 중심형, 여가생활(문화) 중심형, 가정 중심형, 식료품 중심형 소비패턴에서 여성의 비중이 높았다.
- 가구주 연령은 교육비 중심형, 여가생활(문화) 중심형, 여가생활(여행) 중심형, 복합형 소비패턴에서 낮았다.
- 가구주 학력은 소득이 높은 가구들의 소비패턴에서 높았다. 이는 자동차 보유대수와 주택소유 가능성과도 맥을 같이 하였다.
- 가구원수는 교통비 중심형, 교육비 중심형, 가정 중심형, 보건 중심형, 식료품 중심형 소비패턴에서 많을 가능성이 높았다.

마지막으로 본 연구에서 제시한 소비패턴이 가구의 다양성을 충분히 반영하였는지를 확인하기 위해 가구 유형에 따른 소비 포트폴리오를 살펴 보았다. 그 결과, 기존에 널리 사용되던 소득수준을 기준으로 가구를 유형화하는 방식보다 소비패턴을 더욱 잘 구분시키는 군집을 발견하였음을 확인하였다.

본 연구는 가구의 다양성을 반영한 소비패턴을 제시하였는데, 간단하면서도 성능이 좋은 군집화 모형을 통계청 가계동향조사 데이터에 적합시켜 활용성을 높였다는 점에서 의의가 있다. 본 연구에서 제시한 소비패턴들은 국내 가구의 소비활동의 유형으로 사용될 수 있다. 혹은 본 연구의 군집화 프레임워크를 개인 소비 데이터에

적합시켜 가계가 아닌 개인 소비자의 다양성을 반영한 소비패턴을 찾아내는 데 활용될 수 있다.

그러나 본 연구는 몇 가지 한계를 지닌다. 첫째, 가계동향조사 데이터를 이용한 패널 분석은 불가했기 때문에 소비패턴의 시간에 따른 변화를 분석하지 못 했다. 시간에 따라 변화하는 소비 성향을 잡아내기 위해 소비자의 과거 행동을 기반으로 소비 행동을 정의할 필요가 있기 때문에 후속 연구에서는 패널 분석이 가능한 자료를 이용해 군집 간 이동을 설명할 수 있다. 예를 들어 마르코프 행렬을 이용해 각 군집이 다른 군집으로 얼마의 확률로 전이될 것인가를 파악할 수 있을 것이다. 둘째, 해석의 용이성을 위해 군집 개수를 9개로 설정했지만, 국내 전체 가구의 소비활동을 설명하기에 9개 군집은 부족할 수 있다. 따라서 후속 연구에서는 국내 가구를 연령 등으로 먼저 분류한 후 각 연령 범위 내에서 군집화를 수행할 수 있을 것이다. 이를 통해 가구의 더욱 큰 다양성이 포착될 수 있다. 셋째, 본 연구에서 활용한 계층적 군집화 기법은 데이터에 계층적 구조가 존재할 때 잘 작동하는데, 데이터의 형태에 따라 분할 기반 군집화, 분포 기반 군집화, 밀도 기반 군집화 등 다른 군집화 기법이 더욱 잘 작동할 수 있다. 따라서 더 많은 다양한 군집화 알고리즘으로 실험한 후 모델의 성능과 결과를 비교해볼 필요가 있다.

참고문헌

김인철(2004). 소비- 경제를 움직이는 힘, KDI 나라경제, https://eiec.kdi.re.kr/publish/naraView.do?cidx=4511

성영애. (2013). 군집분석을 통해 살펴본1 인 가구의 연령대별 소비지출패턴. 소비자학연구, 24(3), 157-182.

손상희. (1993). 가계소비패턴의 구조. 소비자학연구, 4(2), 51-72.

정영숙. (2000). 소비지출패턴: 연구동향과 미래전망. 소비자학연구, 11(2), 85-101.

정원오, & 이선정. (2011). 빈곤계층의 소비패턴에 관한 연구: 2007 년과2008 년의 변화 비교. 사회복지연구, 42(1), 305-331.

최옥금. (2011). 노인 가구의 소비지출 유형화 및 영향 요인 분석. 노인복지연구, 51, 277-296.

최지혜, 정예림, 박선주, & 박태준. (2021). 코로나19 가 가져온 오프라인 소비 패턴의 변화: 2020 년 신용카드 데이터 분석. 연세경영연구, 58(3), 83-102

Baker, Scott R., et al. "How does household spending respond to an epidemic? Consumption during the 2020 COVID-19 pandemic." The Review of Asset Pricing Studies 10.4 (2020): 834-862.

Chai, A., Rohde, N., & Silber, J. (2015). Measuring the diversity of household spending patterns. Journal of Economic Surveys, 29(3), 423-440.

Chen, C., Petterson, S., Phillips, R., Bazemore, A., & Mullan, F. (2014). Spending patterns in region of residency training and subsequent expenditures for care provided by practicing physicians for Medicare beneficiaries. Jama, 312(22), 2385-2393.

Chronopoulos, D. K., Lukas, M., & Wilson, J. O. (2020). Consumer spending responses to the COVID-19 pandemic: an assessment of Great Britain. Available at SSRN 3586723

Davis, M. A., Nallamothu, B. K., Banerjee, M., & Bynum, J. P. (2016). Identification of four unique spending patterns among older adults in the last year of life challenges standard assumptions. Health Affairs, 35(7), 1316-1323.

Douglas, N., & Douglas, N. (2004). Cruise ship passenger spending patterns in Pacific island ports. International Journal of Tourism Research, 6(4), 251-261.

Estivill-Castro V (2002) Why somany clustering algorithms: a position paper. ACMSIGKDD Explor Newsl 4:65 - 75

Jiang, Z., Zheng, Y., Tan, H., Tang, B., & Zhou, H. (2016). Variational deep embedding: An unsupervised and generative approach to clustering. arXiv preprint arXiv:1611.05148

Ma, Q., Zheng, J., Li, S., & Cottrell, G. W. (2019). Learning representations for time series clustering. Advances in neural information processing systems, 32.

McConville, R., Santos-Rodriguez, R., Piechocki, R. J., & Craddock, I. (2021, January). N2d:(not too) deep clustering via clustering the local manifold of an autoencoded embedding. In 2020 25th International Conference on Pattern Recognition (ICPR) (pp. 5145-5152). IEEE.

McCully, C. P. (2011). Trends in consumer spending and personal saving, 1959-2009. Survey of Current Business, 91(6), 14-23.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and

validation of cluster analysis. Journal of computational and applied mathematics, 20, 53-65.
Schanzenbach, D. W., Nunn, R., Bauer, L., & Mumford, M. (2016). Where does all the money go: Shifts in household spending over the past 30 years. Brookings Institution, The Hamilton Project
Shin, S. J., Song, K., & Moon, I. C. (2020, April). Hierarchically clustered representation learning. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 34, No. 04, pp. 5776-5783).
Xie, J., Girshick, R., & Farhadi, A. (2016, June). Unsupervised deep embedding for clustering analysis. In International conference on machine learning (pp. 478-487). PMLR.
Xu, D., & Tian, Y. (2015). A comprehensive survey of clustering algorithms. Annals of Data Science, 2(2), 165-193.
Yang, J., Parikh, D., & Batra, D. (2016). Joint unsupervised learning of deep representations and image clusters. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 5147-5156).
Yocum, C. (2007). Household spending patterns: A comparison of four census regions. US Bureau of Labor Statistics Working Papers, (412).