점포당 매출액을 활용한 서울 소재 외식업종별 상권 분석에 관한 연구

장소라 [1], 황재호 [2], 서수연 [2], 민무홍 [3]
[1] 성균관대학교 데이터사이언스융합학과 석사과정
[2] 성균관대학교 실감미디어공학과 석박통합과정
[3] 성균관대학교 학부대학 조교수

soraaaa95@g.skku.edu, ashiereki@g.skku.edu, sooyon1119@g.skku.edu, iceo@skku.edu

A Study on Market Analysis of Seoul's Commercial Districts by Food Service Sector Using Sales per Store

Sora Jang[1], Jaeho Hwang[2], Sooyon Seo[2], Moohong Min[3]
[1]Dep. of Applied Data Science, Sungkyunkwan University
[2]Dept. of Immersive Media Engineering, Sungkyunkwan University
[3]University College, Sungkyunkwan University

요 약

본 연구는 서울소재 외식업종의 6 년간 점포당 매출액 데이터를 이용해 시계열 군집분석을 수행, 업종 및 지역별 상권을 세분화하고 '성장 상권'부터 '쇠퇴 상권'에 이르기까지 재정의한다. 이를 통해 예비 창업자와 소상공인이 업종과 지역을 선정하는 지표들을 분석하고 연구하였다.

1. 연구 배경

통계청의 2021 년 소상공인 실태조사에 따르면, 소상공인이 직면한 주요 경영 애로 사항으로 경쟁 심화, 원재료비 부담, 상권 쇠퇴 등이 있다. 특히 외식업은 개인 사업체의 비중이 높고 창업비용 대비 매출액이 상대적으로 낮은 특성을 지니고 있다. 한국농수산식품유통공사의 통계에 따르면, 외식 업체의 폐업률 역시 연간 증가 추세에 있다.

이러한 환경에서 상권을 이해하고 분석하는 것은 창업자와 소상공인에게 매우 중요한 문제이다. 상권 분석에는 업종, 인구, 교통, 임대료 등 다양한 변수가 포함되며, 이들을 종합적으로 고려해야 한다[1]. 하지만 그 중에서도 업종별-지역별로 상권이 어떻게 형성되어 있는지를 판단하는 것이 가장 핵심적이다. 지금까지 상권 분석을 위한 다양한 시도가 이루어졌지만 공공데이터를 활용하여 업종과 입지를 동시에 고려한 연구는 많지 않았다. 또한 상권은 단기간에 형성되거나 변화하는 것이 아니므로 비교적 장기적인 시각에서의 분석이 필요한데, 기존 연구들은 주로 1~2 년 내외의 단기간에 한정하여 분석했다는 점에서 한계가 존재한다.

2. 연구 목적

2.1. 업종과 입지를 동시에 고려: 본 연구는 서울시의 공공데이터를 활용하여 서울 소재 외식업종의 행정동별 상권 성장 및 쇠퇴 상황을 판단하고 분석한다.

2.2. 장기적 트래킹: 2017 년부터 2022 년까지의 6 년 시계열 데이터를 분석하여 비교적 장기적인 상권의 변화를 파악한다.

2.3. 실질적인 도움 제공: 단순 매출액이 아닌 점포당 매출액을 지표로 활용하여 외식업 예비 창업자와 소상공인이 업종과 입지 선택에 있어 더욱 합리적인 결정을 할 수 있는 단서를 제공하고자 한다.

3. 관련 연구

3.1. 상권 분석 연구

예비 창업자에게 유용한 정보를 제공하기 위한 상권 분석 연구는 다양한 방식으로 이루어지고 있다. 기존의 연구들은 상권의 성장성을 나타내는 지표로 주로 매출액을 사용하였다.

기존 연구를 살펴보면 서울시의 골목상권 데이터를 활용하여 매출액을 종속 변수로, 인구, 소득, 집객 시

설 등을 설명 변수로 한 다중 회귀 분석을 실시하였다[2]. 이 연구에 따르면 특히 20~40 세의 소득이 있는 인구와 출퇴근 시간대의 매출 비율이 높을수록 상권의 매출액이 증가하는 경향이 있었다.

골목상권의 외식업 평균 매출액에 영향을 미치는 변수를 분석하여 예측 모델을 구축한 연구에서는 랜덤포레스트 모델을 사용하여 업종과 행정구, 평균 영업 개월 수 등이 평균 매출액에 큰 영향을 미치는 것을 확인하였다[3].

군집 분석을 통해 골목상권을 성장 상권과 정체/쇠퇴 상권으로 구분한 뒤 로지스틱 회귀 분석을 수행한 연구에서는 여성 및 20~30 대의 매출 비율, 건축물 밀도, 상권 면적 등이 성장 상권에 긍정적인 영향을 미치는 것을 확인했다[4].

단, 이러한 기존 연구들은 대부분 1~2 년의 단기 데이터를 분석했다는 점, 상권과 업종을 동시에 고려한 분석이 아니었다는 점에서 한계가 있다.

3.2. 시계열 군집 분석 연구

시계열 군집 분석은 유사한 데이터를 라벨 없이 동질적인 집단으로 분류하는 군집 분석의 한 유형으로, 연속적인 시계열 데이터 내의 숨겨진 패턴들을 파악하여 의미있는 정보를 발견하는데 유용하게 활용되고 있다.

한 연구에서는 시계열 데이터에 대한 군집 분석을 위해 동적 시간 왜곡(DTW)을 사용한 퍼지 C-Means, 퍼지 C-Medroid, 그리고 이 두 알고리즘을 혼합한 하이브리드 알고리즘을 제안하였다. 이러한 퍼지 군집 알고리즘은 K-means 와는 달리 각 군집에 속할 확률을 계산하여 더 다양한 정보를 제공한다. 이를 8 개의 다양한 시계열 데이터셋에 적용한 결과, 대부분의 데이터셋에서 유클리디안 거리보다 높은 정밀도를 보였다[5].

시계열 데이터는 차원이 높고 노이즈가 많아 전통적인 군집 알고리즘으로는 한계가 있다[6]. 이를 해결하기 위해 딥 뉴럴 네트워크를 활용한 딥 클러스터링 방법론이 대두되기도 하였다.

오토인코더 기반의 군집화 알고리즘은 독립변수의 내재적 특징을 추출하여 저차원으로 표현한 데이터를 군집화하는 것이 특징인데, 대표적인 방법으로 DEC(Deep Embedding Clustering)가 있다[7]. DEC 는 오토인코더를 통해 저차원으로 압축한 데이터를 군집화하고, KL-Divergence 손실함수를 활용해 중심값을 최적화시킨다. 이러한 DEC 알고리즘을 시계열 데이터에 적용한 연구도 있는데[8], 이 연구에서는 연속적인 시계열 특성을 반영하기 위해 시계열 데이터를 시간창의 형태로 전처리하고 사전학습을 진행하였다. 학습된 오토인코더를 통해 도출된 잠재변수를 바탕으로 K-means 군집분석을 진행한 후, KL-Divergence 를 통해 중심값을 최적화하는 방향으로 학습한다는 점에서 DEC 알고리즘을 시계열 데이터에 적용한 사례임을 알 수 있다.

4. 연구 방법

본 연구에서는 총 10 종의 외식업종에 대한 6 년의 서울시 행정동별 매출액 관련 데이터를 기반으로 시계열 군집분석을 수행하여 업종별로 상권을 세분화하고자 한다. 군집화된 상권을 점포당 매출액의 성장추이와 규모를 토대로 ‘성장 상권’, ‘기대 상권’, ‘전통강호 상권’, ‘정체 상권’, ‘쇠퇴 상권’으로 재정의하여 예비 창업자가 업종을 선택하거나 지역을 선택할 때 실질적인 도움이 될 수 있게 하고자 한다.

4.1. DTW 를 활용한 K-means 군집 알고리즘

K-means 군집 알고리즘은 두 개의 단계로 구성된다. 첫 번째 단계에서는 군집의 개수인 k 를 고정시킨 후, k 개의 중심을 랜덤하게 선택한다. 다음 단계에서는 각 데이터 객체들을 가장 가까운 중심점으로 이동시켜 초기 군집을 설정한다. 형성된 군집에서 군집의 평균값을 통해 중심점을 재계산하고, 변경된 중심점을 기준으로 데이터 객체들을 재할당하는 과정을 클러스터의 중심이 변하지 않을 때까지 반복한다.

각 데이터 객체를 가까운 군집의 중심으로 할당하는 과정에서 객체와 중심점 간 거리를 계산해야 하며, 거리를 측정하는 방식에 따라 군집 결과가 상이하게 나타날 수 있다. 일반적인 거리 측정 지표로는 유클리디안 거리 등이 있다. 시계열 데이터에서 유클리드 거리는 비교하고자 하는 시계열의 각 데이터 포인트들을 동일한 시간선상에 두고 거리를 측정하게 되는데, 만약 유사한 패턴을 보이더라도 각 시계열에서 해당 패턴이 약간의 시간차를 두고 발생한다면 유클리디안 거리는 유사성을 찾기 어려워진다는 단점이 존재한다.

본 연구는 6 년의 시계열 데이터를 기반으로 하는 바, 약간의 시간적 왜곡을 허용하여 두 시계열 비교 시 적절한 위치를 매칭시킬 수 있는 DTW 를 거리 측정 지표로 사용하고자 한다. 유클리디안 거리와 DTW 방식을 각각 적용한 K-means 군집을 수행하여 결과로 도출된 군집 내 응집도를 비교하고자 한다.

4.2. Elbow Method 와 Silhouette Score 를 활용한 군집 개수 결정

업종별로 상권을 구분해 줄 수 있는 최적의 군집 개수를 결정하기 위해 군집 개수에 따른 군집 내 오

차제곱합을 시각화한 Elbow Method 와 Silhouette Score 를 상호보완적으로 활용하고자 한다.

Elbow Method 는 군집의 개수를 결정하기 위한 전통적인 방법 중 하나로, 군집 개수를 하나씩 증가시켜 갈 때마다 학습 비용을 계산함으로써 학습 비용이 가장 급격하게 감소하는 시점의 군집 개수(k)를 최적의 개수로 판단하는 방법이다[9]. 이 때 학습 비용이란 각 군집별 중심점과 해당 군집에 포함되어 있는 데이터 샘플들 간의 거리를 제곱하여 합산한 오차제곱합(Sum of Squared Error, SSE)을 의미하며, 이 오차제곱합을 최소화하는 방향으로 학습이 진행된다. 다만 급격하게 감소하는 시점 없이 오차제곱합이 일정하게 감소하는 경우, elbow 를 발견하기 어려워 군집의 개수를 결정하기가 어렵다는 단점이 존재하기도 한다.

실루엣 계수는 개별적인 샘플과 타 샘플 간의 거리를 계산함으로써 산출되며, 자신이 속한 군집이 타 군집에 비해 얼마나 유사한지를 보여주는 지표이다. 그리고 전체 데이터에 각 계수의 평균을 취한 값을 실루엣 점수(Silhouette Score)라고 한다. 실루엣 점수가 높을수록 군집이 잘 형성되었다고 할 수 있다.

4.3. 상권 정의 기준

데이터를 토대로 군집 분석을 수행한 후, 군집별 특성을 토대로 업종별 상권을 정의할 것이다. 이를 위해 업종별-행정동별 매출액을 점포수로 나누어 산출한 '점포당 매출액'을 상권 정의의 기준으로 삼고자 한다. 매출액이 영업 성과의 외형적 규모를 대변한다면 점포당 매출액은 이에 더불어 점포 효율성을 나타낸다. 따라서 상권의 성장과 창업자가 취할 수 있는 이익을 동시에 고려할 수 있는 '점포당 매출액'을 상권 판단의 기준으로 한다.

우선 2017 년 대비 2022 년의 점포당 매출액 성장을 비교하고, 다음으로 현 시점의 점포당 매출액 규모를 비교하여 '성장 상권', '기대 상권', '전통강호 상권', '정체 상권', '쇠퇴 상권'의 총 5 가지 상권으로 정의하고자 한다. 2017 년 대비 2022 년의 점포당 매출액이 성장하는 추세이고 타 군집대비 점포당 매출액의 절대적 규모 또한 크다면 '성장 상권'이다. 점포당 매출액이 성장 추세이지만 그 규모가 충분히 크지 않은 경우는 '기대 상권'으로 정의한다. 반대로 점포당 매출액은 하락 추세이나 규모가 큰 경우는 '전통강호 상권'으로, 점포당 매출액이 하락 추세이면서 규모 또한 작다면 '쇠퇴 상권'으로 정의한다. 성장 또는 하락이 뚜렷하지 않고 점포당 매출액의 규모도 타 군집과 비교했을 때 중간 정도의 규모라면 '정체 상권'으로 정의하고자 한다.

5. 실험 및 결과

본 연구는 '서울시 상권분석서비스' 및 '서울 열린데이터 광장'에서 수집한 서울시 지역별 현황 데이터 및 상권별 추정매출 데이터를 기반으로 한다. 해당 데이터는 2017 년 1 분기부터 2022 년 4 분기까지 총 24 분기동안 골목상권 외식업종 10 종에 대해 서울시 내 424 개 행정동에서 나타나는 변화를 반영하고 있다. 업종 및 지역 특성관련 변수는 점포수, 신생기업 생존율, 평균 영업 기간, 인구수, 개폐업수, 임대시세, 집객시설 수 등이 있으며, 매출액 관련 변수로는 매출액, 연령별/남녀별/주중/주말 매출비율 등이 있다.

군집 분석 시 상권의 성장 및 하락을 나타내는 매출액 관련 변수를 활용했으며 결측치가 있는 데이터는 제거하였다. 또한 군집 분석 시 사용하는 K-means 등의 알고리즘이 거리를 기반으로 동작하기 때문에 변수 단위로 인한 영향력을 낮춰주기 위해 Standard Scaler 를 적용하였다. 이 외 업종 및 지역 특성 관련 변수는 점포당 매출액을 계산하고 상권에 영향을 미치는 요인을 추가적으로 파악하기 위한 향후 연구 개선 목적을 위해 수집하였으며, 결측치의 경우 행정동보다 상위 행정구역인 자치구의 데이터로 대체하였다.

5.1. 군집 분석 알고리즘 평가

2017 년 1 분기부터 2022 년 4 분기의 점포당 매출액을 입력값으로 하여 K-means, DTW K-means 알고리즘으로 각각 군집 분석을 수행해보고 그 결과를 비교하였다. 레이블이 없는 데이터셋이므로 군집의 정확한 성능 평가가 불가하므로, 군집 내 응집력을 나타내는 오차제곱합과 실루엣 점수를 비교하여 적합한 알고리즘과 군집 개수를 설정하였다. 군집 개수는 해석 상의 문제로 2 개부터 6 개까지를 분석에 사용하였다.

먼저 일반적인 K-means 군집 분석에서는 군집 개수가 2 일 때 업종 평균적으로 군집 내 오차제곱합이 3883.4 에 이르지만, DTW 거리를 활용하여 데이터의 시계열적 특성을 반영한 K-means 에서는 업종 평균 6.128 로 급감하게 된다. 실루엣 점수에서는 두 알고리즘 간 큰 차이가 나타나지는 않지만, DTW K-means 알고리즘을 적용했을 때 실험에 사용한 모든 군집 개수에 대해 보다 점수가 도출되었다는 점에서 DTW K-means 가 실험 데이터에 더 적합하다는 것을 확인할 수 있다.

<표 1> 각 알고리즘으로 도출된 오차제곱합 업종 평균치

오차제곱합	K=2	K=3	K=4	K=5	K=6
K-means	3883.4	3440.7	3151.0	2941.5	2808.0
DTW K-means	6.128	5.347	4.952	4.648	4.458

<표 2> 각 알고리즘으로 도출된 실루엣점수 업종 평균치

실루엣점수	K=2	K=3	K=4	K=5	K=6
K-means	0.326	0.248	0.189	0.150	0.132
DTW K-means	0.341	0.254	0.196	0.162	0.137

5.2. 업종별 상권 군집 결과

위 표 2 에서 확인할 수 있듯이, 군집의 수가 2 개일 때 실루엣 점수는 가장 높게 나타나는 경향이 있다. 군집 개수가 2 인 경우에 생성된 군집은 모든 업종에서 각각 점포당 매출액이 상승하는 상권과 하락하는 상권으로 구분했다는 점에서 대체로 뚜렷한 군집 특성을 보이기 때문이다. 반면 이 경우 군집 내 분산은 가장 크게 나타나므로 응집력이 낮다고도 해석할 수 있을 것이다. 또한 실제로는 점포당 매출액의 상승폭 또는 규모가 크지 않거나 정체 중인 상권이 있다는 점을 고려하여, 세부적인 상권 정의를 위해 군집 내 분산은 다소 감소시키면서도 실루엣 점수를 크게 낮추지 않는 선에서 최종적인 군집 개수를 설정하였다.

그 결과 점포수가 가장 많아 상대적으로 다양하게 상권 정의가 가능한 '한식음식점' 업종은 4 개 군집으로, 다른 업종은 3 개 군집으로 상권을 구분하였으며, 미리 정의한 기준에 따라 업종별로 표 3 과 같이 상권을 재정의하였다. 그와 함께 표 4 와 같이 해당 상권에 속하는 행정동을 구분하였다.

<표 3> 업종별 상권 군집 정의

한식음식점	중식음식점	일식음식점	양식음식점	제과점
성장상권 기대상권 전통강호상권 쇠퇴상권	성장상권 정체상권 쇠퇴상권	성장상권 기대상권 쇠퇴상권	성장상권 정체상권 쇠퇴상권	성장상권 기대상권 쇠퇴상권
패스트푸드	**치킨전문점**	**분식전문점**	**호프간이주점**	**커피음료**
성장상권 정체상권 쇠퇴상권	성장상권 정체상권 쇠퇴상권	기대상권 전통강호상권 쇠퇴상권	기대상권 전통강호상권 정체상권	기대상권 전통강호상권 쇠퇴상권

<표 4> 상권별 대표 행정동(커피음료 업종의 경우)

구분	커피음료		
	기대상권	전통강호상권	쇠퇴상권
대표 행정동	가회동 성수 2 가 1 동 을지로동 외	오륜동 목 5 동 잠실 6 동 외	대치 1 동 중계본동 창 1 동 외

6. 결론 및 논의

본 연구에서는 서울 지역의 외식업종 상권 분석에 필요한 공공데이터를 수집하여 군집 분석을 함으로써 업종별-행정동별 성장 중인 상권과 그렇지 않은 상권을 판단할 수 있는 유의미한 결과를 도출할 수 있었다. 특히 단순 매출액이 아닌 점포당 매출액을 사용하는 것이 창업자의 입장에서 점포 효율성이 우수한 상권을 판단하는데 보다 실질적인 지표가 될 수 있음을 확인하였다.

하지만 아직은 수집할 수 있는 데이터의 양과 질이 충분하지 않다는 점, 실험 목적에 보다 적합한 군집 분석 방법론이 있을 수 있다는 점, 업종별-행정동별 정의된 상권에 영향을 미친 요인으로 어떤 것이 있는지를 세부적으로 분석하지 못한 점 등은 향후 개선의 여지가 있는 부분으로 판단된다.

본 연구는 6 년의 시계열 데이터를 기반으로 하였으므로 보다 장기적인 상권의 흐름을 분석했다는 점에 의의가 있다. 그리고 업종과 입지를 동시에 고려하였기에 특정 업종을 창업하려는 사람은 어느 지역이 유리할지, 반대로 특정 지역에서 창업하려는 사람은 어느 업종이 유리할지를 판단할 수 있는 근거를 제공하는 연구를 진행하였다.

본 과제(결과물)는 교육부와 한국연구재단의 재원으로 지원을 받아 수행된 첨단분야 혁신융합대학사업의 연구결과입니다.

참고문헌

[1] 이의종, "창업자를 위한 부동산 입지 및 상권분석에 관한 연구", 한국전자통신학회지, 제 3 권, 제 1 호, 67-73, 2009

[2] 김현철, 이승일, "서울시 골목상권 매출액에 영향을 미치는 요인에 관한 연구", 서울도시연구, 제 20 권, 제 1 호, 117-134, 2019

[3] 임소연, "골목상권 내 외식업종 점포의 월 매출액 예측 모형에 관한 연구", 이화여자대학교, 2018

[4] 강현모, 이상경, "시계열 군집분석과 로지스틱 회귀분석을 이용한 골목상권 성장요인 연구", 한국측량학회지, 제 37 권, 제 6 호, 535-543, 2019

[5] Hesam Izakian, Witold Pedrycz, Iqbal Jamal, "Fuzzy clustering of time series data using dynamic time warping distance", Engineering Applications of Artificial Intelligence, Volume 39, 235-244, 2015

[6] Ali Alqahtani, Mohammed Ali, Xianghua Xie, Mark W. Jones, "Deep Time-Series Clustering: A Review", Electronics, 10(23), 3001, 2021

[7] Junyuan Xie, Ross Girshick, Ali Farhadi, "Unsupervised Deep Embedding for Clustering Analysis", Proceedings of The 33rd International Conference on Machine Learning, PMLR 48, 478-487, 2016

[8] 윤동희, 김수민, 김도현, "딥러닝을 이용한 시계열 데이터 군집화", 신뢰성응용연구, 제 19 권, 제 2 호, 167-178, 2019

[9] Trupti M. Kodinariya, Dr. Prashant R. Makwana, "Review on determining number of Cluster in K-Means Clustering", International Journal of Advance Research in Computer Science and Management Studies, Volume 1, Issue 6, 90-95, 2013