

# 빅데이터 활용 소상공인 창업지원 점포 분석 사례 연구

김진철\*, 양현철\*\*

\*한국정보화진흥원 K-ICT 빅데이터센터

\*\*미래창조과학부 융합신사업과

mail:cckim@nia.or.kr, yhc98@msip.go.kr

## Case Study on Big Data Analysis Based Store Evaluation for The Startup of Small Traders and Enterprisers

Chin-Chol Kim\*, Hyun-chul Yang\*\*

\*National Information Society Agency

\*\*Ministry of Science, ICT and Future Planning

### 요 약

본 논문에서는 소상공인의 창업 성공을 지원하는 점포 평가 분석 사례를 소개하여 기업의 빅데이터 도입 및 활용을 촉진하고자 한다. 본 사례에서는 카드사 거래 정보, 가맹점 정보, 부동산 가격 정보, 부동산 통계 정보, 감정평가 정보, 조사업무관련 정보 및 인허가 개폐업 정보를 활용해 36만개의 GIS 블록과 GEO 콘텐츠를 생산하여 빅데이터 분석을 실시하였다. 체계적인 분석을 위해 상권 평가 지수, 업종 평가 지수, 입지 평가 지수, 임대료 추정, 매출 추정, 적정면적 추정 등의 상권, 업종, 입지에 대한 지표를 개발하였다. 이를 통해 상가와 상권에 대한 분석 자료를 제공하여 과밀창업의 예방과 신중한 창업의 유도를 통해 창업실패로 유발 될 수 있는 경제적 비용의 감소 효과를 이룰 것으로 판단된다.

### 1. 서론

소상공인 업계는 2012년 기준 통계로 265만 사업자와 직접 종사자수만 720만명에 이르는 방대한 산업계를 이루고 있으며 베이비부머 세대의 본격적인 은퇴에 따라 점차 그 수는 증가세에 있다. 반면 업계에 대한 진입 통제나 안전장치는 여전히 미비하고 자영업자간의 경쟁, 대기업과의 경쟁에서 치열한 생태계를 이루고 있는 상황이다[1,2]. 자영업 창업시장이 어려운 데는 여러 이유가 있었지만 경쟁이 되는 대기업에 비해 진입과 전업을 판단할 정보의 불균형 또한 한 몫을 하고 있다. 소상공인 업계의 근본적인 경쟁력 회복방안은 취업을 통한 과도한 경쟁예방이 보다 근본적인 처방일 수는 있겠지만, 창업을 시도할 때 보다 과학적이고 객관적인 정보에 접근하게 함으로써 준비된 창업이 될 수 있도록 하고 창업 시 위험률을 사전에 감지하게 하여 과잉 창업 예방과 성공률을 높이는 방안을 같이 제시하는 것도 큰 의미가 있는 것으로 진단되고 있다.

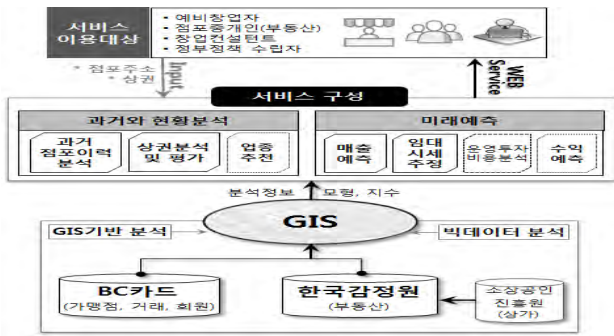
현재 취업자를 포함하여 전 국민 모두가 잠재적인 창업 예비자로 보는 것이 타당하다[3,4]. 반면 창업을 판단하기 위한 객관적이고 유의미한 통계와 분석정보는 미국, 일본, 유럽 등에 비해 대단히 부족하고 열악한 상황이다. 한편 입·출점 판단을 위한 데이터는 예전부터 기업, 공공기관에 존재해왔으나 이의 활용에 대한 법적, 제도적 제약으로 인해 그 활용은 많은 한계를 가졌으나 새로운 비즈니스 기회의 발판을 위해 민간 기업을 중심으로 점차 데이터가 공개되는 추세이고, 국토해양부 공간정보유통체계사업을 비롯해서 공공기관도 새로운 부가가치 창출을 위해 점차 데이터를 개방하는 추세이다[5].

이러한 시장 환경의 변화에 따라 일반인들도 창업의사결정에 도움이 될 만한 진화된 정보서비스의 수요가 늘어나고 있으며, 이런 정보서비스를 통해 정보 불평등에서 오는 소상공인 창업시장의 불합리성을 해소할 필요성이 있다고 판단된다. 즉, 점포의 변화이력과 각 변화단계별로 매출의 변화이력을 추적해 보여주는 것만으로도 점포의 적정한 업종이 무엇일지를 유추하거나 창업을 고민하는 업종이 현 위치에서 적합한지에 대해 판단하는데 많은 도움을 줄 수 있을 것이다[6].

본 논문에서는 빅데이터를 활용하여 소상공인의 창업 및 출점 판단에 유용한 점포입지정보 분석플랫폼의 개발과 창업자가 점포를 분석할 때 점포의 변화이력을 간편하고 신속하게 파악하여 창업 및 업종선정의 의사결정에 필요한 고급 정보를 제공한 사례를 소개한다.

### 2. 요구사항 및 설계

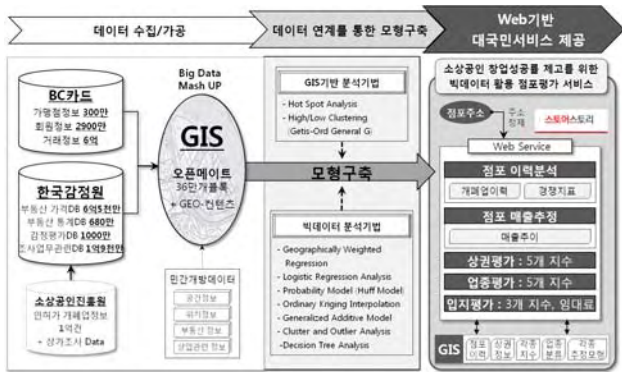
본 논문에서 제안하는 상권분석 시스템은 사용자가 점포의 주소 또는 상권 지역을 입력하면 해당 점포와 상권의 과거와 현황에 대한 분석을 통해 점포이력, 상권평가, 추천업종 등에 대한 정보를 제공한다. 또한, 이후 창업 시에 해당점포에서 발생 할 수 있는 매출, 임대시세, 운영투자비용, 예측수익에 대한 정보 등도 예측한다. 서비스 이용대상은 창업을 준비하며 점포를 물색중인 예비창업자, 상가 위주의 점포 중개인(부동산), 소상공인 대상 전문 창업컨설턴트, 프랜차이즈 가맹본부, 정부의 관련정책 수립자 등이 해당된다.



(그림 1) Summary of Business District Analysis Service

2.1 요구사항 정의

서비스 세부 구성은 아래 그림과 같다.



(그림 2) Detailed Diagram of Business District Analysis

\* **데이터 수집 가공** : 상권 분석에 활용된 빅데이터는 BC카드사의 가맹점정보 300만건과 회원정보 2,900만건, 거래정보 6억건 등 데이터가 활용되었고 한국감정원의 부동산 가격 DB 6억 5천만건, 부동산 통계 DB 680만건, 감정평가 DB 1,000만건, 조사업무관련 DB 1억9천만건 등이 활용된다. 또한, 소상공인진흥원의 인허가 개폐업 정보 1억건과 상가조사 데이터 등이 활용되었다. 이러한 데이터들과 공간정보, 위치정보, 부동산 정보 및 사업관련 정보 등 민간 개방데이터와 매쉬업을 통하여 36만개 GIS 블록과 GEO 컨텐츠 등을 생산하여 분석에 활용한다. 데이터 활용 시에는 상가업소의 매출정보는 창업자에게 매우 중요한 정보이나 개별업소의 매출을 개인정보 및 세원 노출에 대한 법적규제로 인해 제공되지 못하고 있는 상황으로 이에따라서 개별업소에 대한 매출을 보정한(현금비율, 타사카드 비율 반영) 추정값으로 산정한 뒤 지역별, 업종별 유형화, 업종 그루핑(업종 분류), 지역단위별 5개 이하 업소매출 삭제 등 가명처리, 총계처리, 범주화, 삭제, 마스킹 등 비식별화 기술을 적용하여 통계화 한 뒤 정보 서비스에 활용하였다 [7][8].

\* **데이터 연계를 통한 모형구축** : 분석 모형 구축을 위해서는 크게 Hot Spot 분석, High/low clustering 등 GIS기반 분석기법과 Geographically Weighted Regression, Logistic Regression Analysis, Probability Model(Huff Model), Ordinary Kriging Interpolation, Generalized Additive Model, Cluster and outlier Analysis, Decision Tree Analysis 등 빅데이터 분석

기법을 활용하였다.

\* **서비스 제공** : 최종 분석결과로는 소상공인 창업성공률을 제고하는 빅데이터 활용 점포평가 서비스를 제공한다. 웹기반으로 점포이력 분석, 점포 매출추정, 상권평가, 업종평가, 입지평가 등을 제공한다.

2.2 분석 DB설계

공간정보, 카드정보, 부동산 정보, 기타 개방 데이터를 수집하여 블록정보, 상권정보, 주요상권, 업종 코드 표준화 등을 추출하여 분석에 활용할 DB를 구축하였다.

< 표 1 > Process of Populating Database

데이터 확보 및 수집	공간정보 (36만 블록, 주거, 유동인구, 구매력 등 6종)
	카드정보 (가맹점, 고객특성, 매출, 거래정보 등 4종)
	부동산정보 (건물DB, 상가임대사례조사, 인허가업소 DB 등 3종)
표준화 및 GIS 기반 매쉬업	기타 개방 데이터 (아파트, 공시지가, 주요/집객시설, 교통시설 등 22종)
	업종코드 표준화  Geo-Coding <sup>1)</sup> 및 데이터 융합
DB 설계 및 구축	분석 DB 구축 서비스 DB 구축

3. 분석 서비스 모형 개발

모형 개발 부문은 점포 이력 데이터 매칭 모델과 상권 및 업종 평가 지수 모델로 구분 된다.

1) 점포 이력 데이터 매칭

점포 이력 데이터 매칭에서 ‘인허가 업소 변경 및 개/폐업 History’를 기준 데이터로 사용하고 ‘비씨카드 가맹점 데이터’를 비교 데이터로 사용하였다. 매칭은 Geo-coding 결과값이 정좌표인 대상에 적용하였고, 매칭 기준을 지번 주소와 건물의 업소명 유사도가 85% 이상인 경우이고 지번 주소의 주변지가 일치하고 업소명이 포함 관계인 경우로 설정하였다. 분석 결과로 총 163,525건(서울지역 기준)의 매칭 건수가 도출되었으며 매칭율이 54.74%에 달했다.

2) 상권 및 업종 평가 지수

a) 상권 평가 지수

상권 평가 지수의 세부 지수로 성장성 지수, 안정성 지수, 상권 활성화 지수, 고객 구매력 지수, 유동성 지수를 개발하였다. 성장성 지수는 “상권의 성장성을 판단하는 것으로 카드거래건수의 추이(최근 1년간 전년 동월 대비 증감율의 월별 기하평균)를 이용하여 상권내 가맹점의 평균 구매빈도(카드건수)의 증감율을 지수화한 정보”로 정의하였다. 안전성 지수는 “상권내 가맹점의 평균 영업 개월수(업력)을 지수화한 정보”로, 상권 활성화 지수는

“카드거래건수를 이용하여 거래횟수가 많아질수록 상권이 활성화된 것으로 보고 지수화한 정보”로, 고객 구매력 지수는 “상권내 가맹점에서 3번 이상 구매한 고객의 카드 거래총액을 집계하여 인당 평균 카드 거래총액을 지수화한 정보”로, 유동성 지수는 “도보로 이동하는 사람의 수를 도보 가능한 워킹 포인트 및 블록별 지수화한 콘텐츠”로 정의하였으며 36만 블록(소상권) 단위로 구축하였다.

**b) 업종 평가 지수**

업종 평가 지수의 세부 지수로 성장성 지수, 안정성 지수, 활성화 지수, 고객 구매력 지수, 경쟁 밀집도 지수를 개발하였다. 경쟁 밀집도 지수는 “상권내 동종 업종의 가맹점 수를 지수화한 콘텐츠”로 정의하였고 그 외 지수는 상권 평가 지수의 세부 지수와 동일하게 정의하였으며 36만 블록(소상권) 및 35개 중분류 업종 단위로 구축하였다.

**c) 입지 평가 지수**

입지 평가 지수의 세부 지수로 접근성 지수, 가시성 지수, 도보 통행량 지수, 건물 매력도를 개발하였다. 접근성 지수는 “사람들의 접근을 용이하게 하는 주요 대중교통 시설과 해당 건물이 위치한 토지 형태를 파악하여, 해당 건물에 얼마나 쉽게 접근할 수 있는가를 100점 만점기준으로 나타낸 지수”로 정의하였다. 가시성 지수는 “건물이 위치한 대지의 형상, 도로접면 형태와 도로상에서 해당 건물에 대한 가시성을 파악하여 사람들이 얼마나 쉽게 건물을 인지할 수 있는 가를 100점 만점기준으로 나타낸 지수”로, 도보 통행량 지수는 “사람통행을 유발시키는 시설물에 대하여 각각의 시설들이 통행을 유발시키는 크기를 반영하여 해당 건물 앞을 지나는 유동인구의 통행량을 추정하고, 이를 100점 만점기준으로 나타낸 지수”로, 건물 매력도는 “건물 노후도, 엘리베이터 유무, 주차장 유무에 의해 입점 시 시설비용, 고객 편의 측면에서 평가한 건물의 매력도”로 정의하였으며 서울시 업무/상가용 건물을 대상으로 구축하였다. 입지 평가 지수를 통해 접근성이 높은 건물로 평가된 예는 역삼동에 위치한 창립빌딩이었고, 가시성이 높은 건물로 평가된 예는 대치동 포르쉐타워이며, 통행량이 높은 건물의 예로는 명동입구 상가가 선정되었다.

**d) 임대료 추정**

임대료 추정은 2013년 1분기에서 3분기까지의 3,000개 표본 건물 단위의 전체 임대 점포 조사 데이터를 의미하는 한국 감정원 상가임대사례 조사 데이터와 소상공인 진흥원 상담 자료 중 임대시세 기록 데이터를 활용하여 주요 상권/시군구의 1층 기준 평균 임대료를 산정하고 유동인구 지수와 상권 활성도를 활용하여

활성화/비활성화 지역을 군집화 한 후 층별(2층 이상, 지하층) 비율을 산출하고 주요 상권/시군구별 보증금/임대료 산출 과정을 거쳐 주요 상권별 임대료 추정 정보를 도출하는 과정을 통해 추정되었다. 임대료 추정에 대한 결과는 0.76의 상관 계수가 도출되었다.

임대료 추정 절차에서 평균 임대료 산정 시에 상권 내 임대료 조사정보가 10건 이상인 경우에는 핵심상권 영역 평균 임대료를 산정 기준으로 사용하였다. Km 버퍼 영역 내 임대료 조사정보가 20건 이상인 경우에는 핵심상권 1km 버퍼 영역의 평균 임대료를, 동일 시군구 핵심상권 영역 내 임대료 조사정보가 20건 이상인 경우에는 동일 시군구 핵심상권의 평균 임대료를, 시군구 내 임대료 조사정보가 20건 이상인 경우 시군구 평균 임대료를, 광역 시도 내 임대료 조사정보가 20건 이상인 경우에는 광역 시도 평균 임대료를, 전국 임대료 조사정보가 20건 이상인 경우에는 전국 평균 임대료를 산정 기준으로 사용하였다.

임대료 추정 절차에서 활성 지역의 구분 및 층별 비율 산출 시에 시군구 내 레코드가 20건 이상이면 시군구 평균을 산정 기준으로 사용하였다. 광역 시도 내 레코드가 20건 이상이면 광역 시도 평균을, 전국 임대료 조사정보가 20건 이상이면 전국 평균을 산정 기준으로 사용하였으며 활성지역 구분 시 산출 결과의 비율값이 1.17에서 1.3 사이일 경우 활성지역으로 1보다 작을 경우 비활성 지역으로 구분하였고 층 구분 시에는 1층 대비 0.75 보다 작을 경우 지하 1층으로 1층 대비 0.85 보다 작을 경우 지상 2층으로 구분하였다. 임대료 산출 결과 평균 임대료 상승이 가장 큰 지역은 서울시 중구 명동거리 일대로 파악되었고 평균 임대료 하락이 가장 큰 지역은 서울특별시 성북구 석관동 한국예종 근처로 파악되었다.

**e) 매출 추정**

매출 추정 모델링은 전국의 음식/서비스/판매업 35개 중분류 업종을 대상으로 실시하였고 대상 점포 정의, 매출 데이터 정의, 분석 변수 정의, 모델링 결과 반영의 모델링 프로세스 절차에 따라 수행하였다. 대상 점포는 “6개월 이상의 실적을 가진 표준 업종 코드 부여 가맹점”으로 정의하였다. 매출 데이터 정의는 “BC 가맹점 매출액과 상권특성/업종/카드사별 결제 비율 및 현금비율을 고려하여 총 매출액 환산을 통한 종속 변수의 설정”으로, 분석 변수 정의는 “BC 가맹점 정보와 입지 정보, 상권 정보, 시설 정보 등을 고려한 독립 변수의 설정”으로, 모델링 결과 반영은 “사용자 정보 입력 후 적용할 모형을 선택하고 모형에 포함된 변수 집계를 토대로 모형 적용 및 매출액의 산출”로 정의하였고 총매출 추정을 위한 카드 매출액 보정과 이를 이용한 점단위(건물 단위) 매출 추정 모형, 임대료와 추정된 매출액에 대한 적정 면적을 추정하는 모형으로 매출 추정 모델링을 구분하였다.

가맹점 총매출 추정을 위한 카드 매출액 보정

단계에서는 총매출 환산액 보정에서 비씨카드의 평균적인 카드 결제점유율이 20% 정도를 차지하여 정확한 매출액 반영을 위하여 지역별·업종별로 점유율의 보정작업을 진행하였고 카드 점유율 보정에서 약 60,000건의 VAN 데이터를 활용해 월평균 카드 거래액 VAN 데이터와 BC카드 거래액을 가맹점 단위로 매칭하여 지역별, 업종별 평균 카드 점유율을 산정하였다. 또한 현금결제 비율 추정에서 한 점포의 총 매출 정보를 산출하기 위해 POS 데이터를 수급하여 카드매출 대비 현금비율을 산출하였고 현금비율은 동일 업종이라도 상권의 유형에 따라 비율이 다르게 나타나므로 업종별/상권유형별로 현금비율을 산출하여 적용하였다.

**f) 적정면적 추정**

1차 면적은 서비스업종 조사 자료를 이용하여 면적, 임대료 및 매출액이 반영된 선형회귀모형 (시도별 GLM(General Linear Model), 로그-로그모형)을 구축하였으며 BC매출액 추정모형 및 임대료 추정모형에 의해 계산된 매출액 및 임대료를 적정면적 추정모형에 대입하여 1차 면적을 산출하였다.

기준 면적은 시점에 의한 오차를 보정하기 위해 마이크로데이터 사업체부문 자료의 업종별 면적에 대하여 표본조사를 통해 모든 업종에 대해 시도별 구분 없이 백분위 기준 10개 구간을 설정하여 도출하였다. 서비스업종의 경우 사업체부문 자료에 면적이 없어 2005년 사업체총조사 면적을 기준 면적으로 활용하였다.

2차 면적은 1차 면적이 기준 면적의 범위에 맞게 변환되도록 하는 것으로 1차 면적을 아래와 같은 변환식에 대입하여 산출하였으며 매출액 추정모형 구축에 사용된 모든 BC가맹점 데이터에 대해 임대료 및 추정매출액을 산출하고 이를 면적모형에 대입하여 업종별 최댓값, 최솟값, 평균값 및 비율을 계산하였다.

$$\begin{aligned}
 &mean = \text{업종별}(BC\text{업종기준})1\text{차면적의 평균} \\
 &\text{if } 1\text{차면적} < mean, 2\text{차면적} = mean + (1\text{차면적} - mean) * ratio1 \\
 &\text{if } 1\text{차면적} \geq mean, 2\text{차면적} = mean + (1\text{차면적} - mean) * ratio2 \\
 &ratio1 = \frac{(1\text{차면적평균} - \text{기준면적최소값})}{(1\text{차면적평균} - 1\text{차면적최소값})} \\
 &ratio2 = \frac{(1\text{차면적평균} - \text{기준면적최대값})}{(1\text{차면적평균} - 1\text{차면적최대값})}
 \end{aligned}$$

**4. 결론 및 향후 연구계획**

앞으로 상가업소DB는 공유·융합의 단계를 거쳐 마케팅 지원, 지도서비스, 포탈 등과 연계되어 부가적이며 창의적인 비즈니스영역의 개척에 활용 될 것으로 보인다. 상가업소DB와 연계만으로 음식배달서비스, 맛집, 부동산중개, 경매, 상가풍수, 외국인관광, 소셜쇼핑, 골목상권마케팅, 소상공인정책개발과 정책모니터링, 신속한 개폐업 수검으로 포털, 네비게이션의 검색정확도 향상, 클라우드소싱을 통한

3D VR, 스토어 이너뷰 등 다양한 서비스가 가능하다. 나아가 다른 데이터(금융이나 통신)나 서비스와 융합되어 점포실시간건강진단, 매출예측, 점포별 고객개인별 맞춤형 마케팅, 점포중심 로컬마케팅 등 활용 범위가 무궁할 것으로 보인다.

점포평가모형만 놓고 보자면 개인 신용평가, 기업 신용등급 등 개인과 기업에 일반적으로 적용하는 평가모형을 입지중심의 점포에도 시도하였다는데 큰 의의가 있다고 하겠다. 아직 보완해야할 부분이 많지만 상관관점에서 입지에 대해 점수를 부여하기 시작했다는 점에서 상관분석분야에 새로운 전기가 마련되었다고 평가한다. 상권, 입지, 업종에 대해 개별 지표를 개발하고 이를 점수화 했다. 개별점수가 일반화 되지 않아 종합적으로 판단하는 지표로의 활용은 아직 부족하지만 향후 모형고도화를 통해 보다 직관적인 판단기준의 제시가 가능할 것으로 기대한다.

**참고문헌**

[1] 'Big Data Analytics : An assessment of demand for labour and skills', e-skills uk(2013), 2012~2017

[2] 'World wide Big Data Technology and Services: 2012-2015 Forecast', IDC (2011), 2011

[3] SungHyun Kim, 'Big Data Development Trend of Case study in Big Data : A focus of Global Big Data Case)', National Information Society Agency Bigdata Strategy Center, 2015.5, pp. 6-8

[4] 'The Next Frontier for Innovation, Competition, and Productivity', McKinsey Global Institute (2011) Big data

[5] ShinAe Shin, SungHyun Kim, '2013 Bigdata Case Study for realizing creative economy', National Information Society Agency Bigdata Strategy Center, 2014.3, pp. 14-36

[6] Big data Industry Development Strategy for supporting 'Creative Economy and Government 3.0', Ministry of Science, ICT and Future Planning(Related Ministry Union), 2013.12, pp. 6-8

[7] ChinChol Kim, ShinAe Shin, BaeHyun Kim, 'Privacy Data de-identification Case Study in Big Data', National Information Society Agency Bigdata Strategy Center, 2014.5, pp. 40-43

[8] ChinChol Kim, HyunChol Yang, BaeHyun Kim, 'Privacy Data de-identification Technical Use Guide for Big Data', National Information Society Agency Bigdata Strategy Center, 2015.6, pp. 36-44