공간과사회 2016년 제26권 1호(통권 55호): 336~358
http://dx.doi.org/10.19097/kaser.2016.26.1.336

일반논문

빅데이터 분석을 활용한 도시통계 표현 연구
체르노프 얼굴을 활용한 우리나라 광역자치단체의 지역사회건강지표의 표현을 중심으로*

A Big Data Analysis of Urban Statistics Expression –Chernoff Face-Based Expression of Local Community Health Index in Korea

기정훈**

체르노프 얼굴(Chernoff Face)은 다차원 통계 데이터를 사람의 얼굴 이미지를 이용해 시각적으로 표현하는 방법이다. 체르노프 얼굴은 다변량 자료의 각 변수를 인간 얼굴의 눈, 코, 입 등의 얼굴 특징에 대응시켜 각 관측치를 각기 다른 얼굴로 표현하는 것으로 도시나 지역의 다양한 데이터를 얼굴로 표현한다는 장점과 공간적 인식의 단순성이라는 장점을 가지고 있어서 빅데이터 분석의 표현기법으로 활용된다.

본 연구에서는 체르노프 얼굴을 활용해서 우리나라 광역자치단체들의 지역사회건강에 대한 국민체감지표를 살펴보았다. 여기에서 다루고자 하는 광역자치단체의 통계는 여덟 가지로 (1) 흡연(남자 현재흡연율), (2) 음주(고위험음주율), (3)신체활동(걷기 실천율), (4) 비만(체중조절시도율), (5) 구강건강(점심식사 후 칫솔질 실천율), (6) 정신건강(우울감 경험률), (7) 질병-1(고혈압 진단자 비율), (8) 질병-2(당뇨병 진단자 비율)가 포함된다. 각 통계자료는 질병안전본부에서 제공하는 2014년도 기준 지역사회건강조사의 자료를 통해서 구축했다.

종합적으로 보았을 때에 특별시와 광역시의 지표들이 우수하게 나타나는 것을 얼굴 모양에서 알 수 있다. 대전광역시, 서울특별시, 대구광역시가 우수한 편이다. 반면에 도 단위 자치단체들의 성적은 다소 저조한데, 특히 강원도, 충청북도, 충청남도, 제주특별자치도에서는 개선을 위한 노력이 필요한 것으로 보인다.

체르노프 얼굴을 활용한 지역분석과 표현기법은 기존의 데이터분석이나 표현을 향상시키는 좋은 예가 될 수 있는데, 이를 지역홍보 및 마케팅에 활용할 수 있고, 지역 주민들이 쉽게 이해할 수 있는 얼굴형식이라는 점에서 선거 전략으로도 발

* 이 연구는 2015년도 명지대학교 교책중점연구소 지원으로 연구되었음.

** 명지대학교 행정학과(johnki@mju.ac.kr)

전시킬 수 있는 가능성이 있다. 그러나 얼굴의 각 부위가 어떤 변수를 표현할 지에 대한 명확한 기준이 부족하다는 점 등은 극복해야 할 한계로 남아 있다.

주요어: 빅데이터 분석, 체르노프 얼굴, 지역사회건강지표

1. 서론

빅데이터(Big Data)란 형식이 매우 다양하고 유통되는 속도가 너무 빨라 기존 방식으로 관리 및 분석하기 어려운 데이터를 의미한다. 스마트 기기와 SNS기 확산되면서 빅데이터가 급증하고 있고, 이용자들이 생성한 데이터의 형식(정형, 비정형) 여부를 떠나 데이터 자체가 증가하자 빅데이터의 저장과 가공, 분석, 표현에 대한 관심이 커지고 있다.

그렇지만 빅데이터는 최근까지 컴퓨터공학, 통계학, 수학 등 이공계열의 전유물로 여겨져서, 인문사회과학과는 관계가 적은 영역으로 간주되는 경향이 있다(기정훈, 2014a). 따라서 도시공간을 다루는 인문사회과학 연구자들은 빅데이터를 자신들과는 연관성이 적은 연구분야로 생각하고, 이를 활용한 연구를 진행하거나 공간적 통찰을 빅데이터와 연결하려는 시도가 적었다. 즉, 이 논문은 빅데이터라는 주제가 다양한 연구분야에서 많은 연구자에게 알려져 있지만, 아직 실제적인 접근은 부족하다는 인식에서 시작된다(기정훈, 2014b). 실제적 측면에서 보면, 오히려 도시공간과 관련된 연구에서 빅데이터가 활용될 가능성이 높은데, 특별히 빅데이터 분석의 시각화를 활용한 인포그래픽(infographic)[1]과 같은 방법론을 통해서 다양한 공간적 표현을 할 수 있기 때문에 그 잠재적 가치와 활용

1) 인포그래픽(infographic)이란 인포메이션(information)과 그래픽(Graphic)의 합성어로 다량의 정보를 차트, 지도, 다이어그램, 로고, 일러스트레이션 등을 활용해 한눈에 파악할 수 있도록 하는 디자인을 의미한다.

도가 매우 높다고 볼 수 있다. 본 연구에서는 인포그래픽과 연계된 빅데이터 표현기법의 하나인 체르노프 얼굴을 소개하면서 빅데이터 표현기법에 접근성을 높이는 시도를 하고자 한다. 빅데이터 분석에 대한 장벽이나 문턱을 낮추기 위해서 상대적으로 쉽고 흥미롭게 만들 수 있는 체르노프 얼굴을 통해 관련분야의 연구자들이 다른 빅데이터 분석까지 연구영역을 넓힐 수 있을 것이다.

체르노프 얼굴에 대한 구체적인 사례로서 우리나라 광역자치단체들의 지역사회건강에 대한 통계지표를 살펴보고자 한다. 본 연구는 공간과 관련된 인문사회과학분야에서 빅데이터가 어떻게 활용될 수 있는지 보여줄 수 있는데, 특히 빅데이터의 다양한 표현기법을 활용하면 공간이나 도시 관련 연구자들이 가진 통찰력을 보여줄 수 있음을 확인할 것이다.

2. 빅데이터 분석 기법으로서의 체르노프 얼굴

1) 빅데이터 분석기법의 개요

(1) 빅데이터의 특징과 구성요소

빅데이터는 서론에서 설명한 대로 데이터 형식이 매우 다양하고 그 유통 속도가 너무 빨라 기존 방식으로 관리하거나 분석하기 어려운 데이터를 의미한다(한국콘텐츠진흥원, 2014). 따라서 빅데이터에는 다섯 가지의 주요한 특징을 가지고 이는 '5V'라고 한다(정지선, 2012). 빅데이터의 첫 번째 특징은 데이터의 규모(Volume)가 크다는 것이다. 빅데이터는 테라바이트[2] 또는 그 이상의 데이터 크기를 의미하며 GPS, 각종 센서, SNS 등에서 생성된 대용량 데이터가 이에 해당한다. 두 번째 특징은 다양성(Variety)인데, 데이터의 크기 및 형식이 다양한 비정형 데이터를 포함하

2) 1테라바이트(TB)는 1024기가바이트(GB)의 저장용량을 의미한다.

기 때문에 붙여진 특징이다. 세 번째 특징은 속도(Velocity)인데, 데이터가 생성되고 유통되고 활용되는 데 소요되는 시간이 짧기 때문이다. 빅데이터의 유통 및 활용 속도를 향상시키는 데는 페이스북이나 트위터와 같은 소셜 네트워크 서비스가 큰 역할을 했다(송길영, 2011). 네 번째 특징은 정확성(Veracity)으로 데이터에 부여할 수 있는 신뢰의 수준이 높아졌다는 의미이다. 다섯 번째 특징은 가치(Value)인데 데이터의 전파성이 빠르고 변화의 패턴을 발견하는데 비용이 든다는 뜻이다.

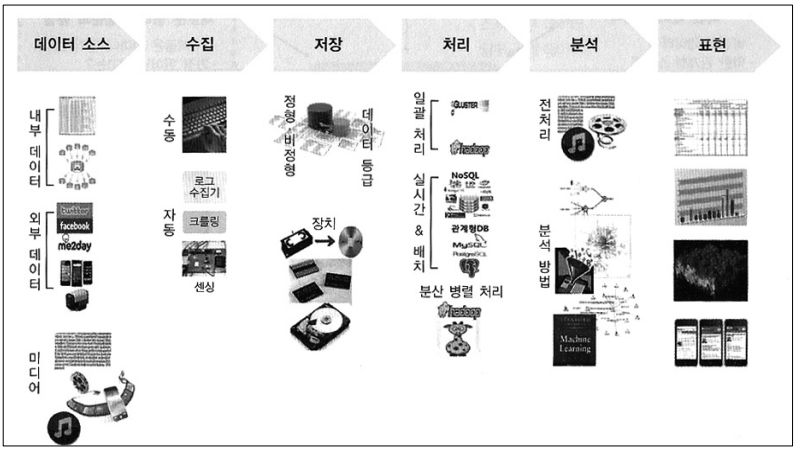
빅데이터의 협의적 정의는 이와 같이 데이터의 소스, 수집, 저장과 같은 물리적 측면과 기술적 측면이 강조되지만, 더 넓은 의미에서 보면 데이터의 분석기법이나 표현기법이 기존의 방식과 다른 것도 빅데이터 분석에 포함시킬 수 있다. 이에 대해서는 '2) 빅데이터의 처리과정과 분석기술'에서 좀 더 자세하게 다루고자 한다. 따라서 빅데이터의 정의는 공학기술과 소셜 네트워크에 기반을 두고 있는 기존의 데이터보다 크고 빠르고 다양한 양상의 데이터뿐 아니라 통계기법의 융합이나 표현기법의 다양화를 포함하는 데이터와 분석기법이라고 할 수 있다.

빅데이터의 구성요소에는 자원, 기술, 인력이 있다(정지선, 2012). 여기에서 말하는 빅데이터의 자원은 데이터 자원 확보와 품질관리에 관한 것이다. 기술은 빅데이터 플랫폼을 의미하는데, 이것은 저장 및 관리, 데이터 처리, 분석 및 시각화를 포함한다. 빅데이터를 분석하고 해석하는 인력을 데이터 사이언티스트라고 한다. 이들은 수학 및 공학 능력을 가지고 있고 경제학, 통계학, 심리학 등 다학제적 이해 능력을 가지고 있어야 한다. 특히 데이터 사이언티스트들은 비판적 시각과 커뮤니케이션 능력을 가지고 있다는 점과 스토리텔링에 능숙하다는 점, 그리고 무엇보다도 시각화를 하는 능력을 가지고 있다.

(2) 빅데이터의 처리과정과 분석기술

빅데이터의 처리과정과 분석에 사용되는 기술을 알아보는 것은 빅데이터를 활용하고 도시공간 분야의 역할을 알아보는 데 중요한 시사점을

<그림 1> 빅데이터 처리과정



자료: 정지선, 2012.

보여준다. <그림 1>와 <표 1>을 보면 빅데이터 처리 과정과 각 과정별로 기술적인 요소들을 볼 수 있는데, 그 과정은 데이터 원자료나 데이터 소스에서 시작해 데이터의 수집, 저장, 처리 과정을 거친다. 그다음에는 처리된 데이터를 분석하고 그 분석결과를 표현하는 과정이 따른다는 것을 알 수 있다.

데이터 소스는 해당 기관이나 조직의 내부 데이터가 될 수도 있고 트위터나 페이스북과 같은 외부 데이터나 영상이나 사진과 같은 미디어가 될 수도 있다. 도시 및 공간분야에서는 지리정보와 관련된 대규모 데이터를 활용하는 사례가 일반적이다. 그다음 단계인 수집단계에서는 검색엔진을 사용해서 인터넷상의 자료를 대량으로 수집하는 '크롤링'이나 사물인터넷과 연관된 센싱을 활용한 데이터 수집은 컴퓨터공학을 중심으로 한 공학적 영역이라고 할 수 있다. 도시 및 공간의 데이터를 모으기 위해서는 지리정보시스템을 활용하거나 CCTV나 도로정보 등을 활용하게 된다. 그다음 단계인 데이터 저장은 공학적 기술이 집약된 단계인데 특히 빅데이터 데이터는 그 규모가 크기 때문에 그 저장용량이 크고 특

<표 1> 빅데이터 처리과정별 기술요소

과정	영역	개요
생성	내부 데이터	데이터베이스Database, 파일 관리 시스템File Management System
	외부 데이터	인터넷으로 연결된 파일, 멀티미디어, 스트림
수집	크롤링Crawling	검색 엔진의 로봇을 사용한 데이터 수집
	ETLExtraction, Transformation, Loading	소스 데이터의 추출 · 전송 · 변환 · 적재
저장	NoSQL 데이터베이스	비정형 데이터 관리
	스토리지Storage	빅데이터 저장
	서버Server	초경량 서버
처리	맵리듀스MapReduce	데이터 추출
	프로세싱Processing	다중 업무 처리
분석	NLPNeuro Linguistic Programming	자연어 처리
	기계 학습Machine Learning	기계 학습으로 데이터의 패턴 발견
	직렬화Serialization	데이터 간의 순서화
표현	가시화Visualization	데이터를 도표나 그래픽적으로 표현
	획득Acquisition	데이터의 획득 및 재해석

자료: P. Warden, 2011.

히 비정형 데이터의 경우는 NoSQL(Not Only SQL)이라는 데이터베이스를 통해서 관리한다. 데이터 저장 후에 시작되는 데이터 처리에 있어서도 맵리듀스(MapReduce)를 통한 데이터 추출이나 하둡(Hadoop)을 통한 분산 병렬처리와 같이 기술적인 측면이 강조된다.

빅데이터를 분석하고 표현할 때에 도시나 공간에 관한 것이 많을 수 있는데, 예를 들면 데이터의 공간적 패턴을 발견하거나 이를 통해 기술적·정책적으로 대응을 하는 사례를 가정할 수 있다. 특히 빅데이터의 표현은 도시와 공간과 관련된 데이터를 분석한 결과를 도표나 그래프로 표현하고 이를 재해석하는 것으로서 본 연구에서 보여주고자 하는 체르노프 얼굴의 기법도 역시 빅데이터 표현기법에 포함된다.

2) 체르노프 얼굴과 활용

(1) 체르노프 얼굴(Chernoff Face)이란?

체르노프 얼굴은 다차원 통계 데이터를 사람의 얼굴 이미지를 이용해서 시각적으로 표현하는 방법이다. 체르노프(Chernoff)가 1973년에 다차원 자료를 이차원 평면상에 표현하기 위해 처음으로 인간의 얼굴을 사용하면서 알려졌다. 체르노프 얼굴은 다변량 자료의 각 변수를 인간 얼굴의 눈, 코, 입 등의 얼굴 특징에 대응시켜 각 관측치를 각기 다른 얼굴로 표현한다. Chernoff(1973)는 자메이카의 에오세 석회암 지층으로부터 고생시대의 87개 원생동물 화석을 관찰해서 배아실에 관련된 6개의 변수에 얼굴 특징을 대응시켜 87개 화석의 얼굴을 이미지에 따라 군집화했다.

체르노프 얼굴은 얼굴의 가로 너비, 세로 높이, 눈, 코, 입, 귀 등 각 부위를 변수로 대체해서 데이터의 속성을 쉽게 파악하기 위해 만들어졌고, 최대 15개의 변수들을 표현할 수 있다. 체르노프 얼굴은 실생활에서 사람의 얼굴을 쉽게 구분한다는 점에서 착안해서 데이터 표현에 따라 달라지는 작은 차이도 쉽게 구분할 수 있다고 가정한다.

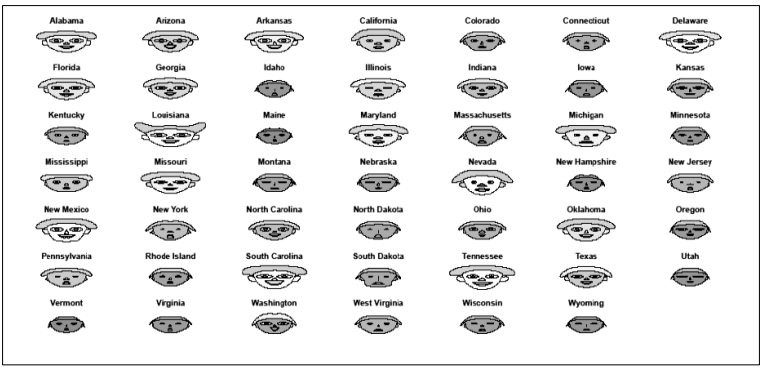
(2) 체르노프 얼굴의 응용 및 활용 사례

미국의 경우에는 <그림 2>에서 보는 것과 같이 각 주(state)별로 강력범죄와 재산범죄의 양상을 체르노프 얼굴로 표현한 사례가 있다. 이를 통해 각 주정부의 담당자는 자신의 주에 맞는 범죄정책을 세우는 데 도움을 받을 수 있고, 유사한 주의 성공사례를 활용할 수도 있다 (Thoplan, 2014).

이와 유사한 분석으로는 미국에서 1976년 이후의 사형집행 통계를 이용한 체르노프 얼굴이 있는데, 여기서는 각 주별로 사형수단, 사형자 중 백인의 비율, 사형자 평균연령, 사형집행의 수와 같은 사형통계를 지도상에 표현했다는 점이 흥미롭다(Huffman, 2010).

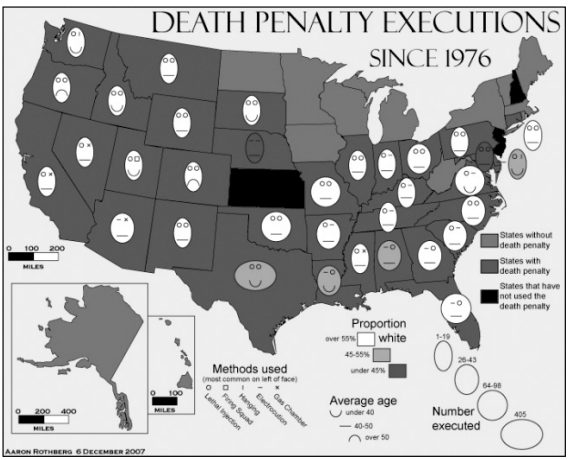
오늘날까지 유명하게 회자되고 있는 체르노프 얼굴의 예는 Turner

<그림 2> 미국의 2012년 주별 범죄양상에 따른 체르노프 얼굴



자료: Thoplan, 2014.

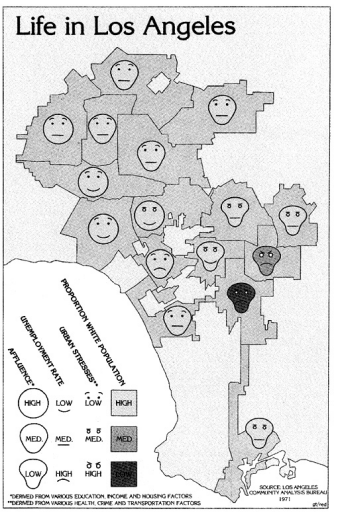
<그림 3> 미국의 1976년 이후 사형집행 양상에 따른 체르노프 얼굴



자료: Huffman, 2010.

(1979)가 만들어 '*Life in Los Angeles 1970*'라는 제목이 붙은 지도다. 이것은 미국 로스앤젤레스 지역의 삶의 질을 나타낸 지표로서 부의 정도, 실업률, 스트레스 지수, 백인의 비율과 같은 4개 변수를 얼굴 모양, 입의 곡률(곡선의 방향), 눈썹 기울기, 그리고 얼굴 색깔과 같은 얼굴 특징에 대응시

<그림 4> 로스엔젤레스의 1970년 생활상에 따른 체르노프 얼굴



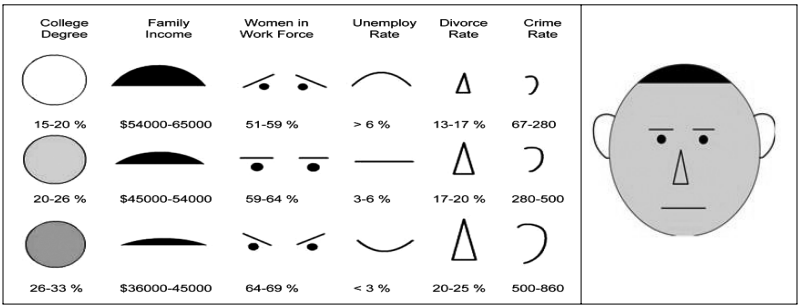
자료: Turner, 1979.

켜 로스앤젤레스의 각 소지역을 체르노프 얼굴로 표현했다. 이 지도에서 행복한 얼굴을 보여주는 지역은 삶의 질이 좋은 지역으로 해석되었다.

Spinelli와 Zhou(2004)는 체르노프 얼굴 분석을 통하여 미국인의 삶의 질을 나타내는 지도를 완성했다. 삶의 질을 결정한다고 보는 여섯 개 변수인 이혼율, 여성노동율, 대학 학위, 범죄율, 실업률, 평균 가구수입에 얼굴의 특징으로서 코의 크기, 눈썹의 방향, 얼굴 색, 귀의 크기, 입의 곡률, 그리고 머리카락의 양을 각각 대응시켰다.

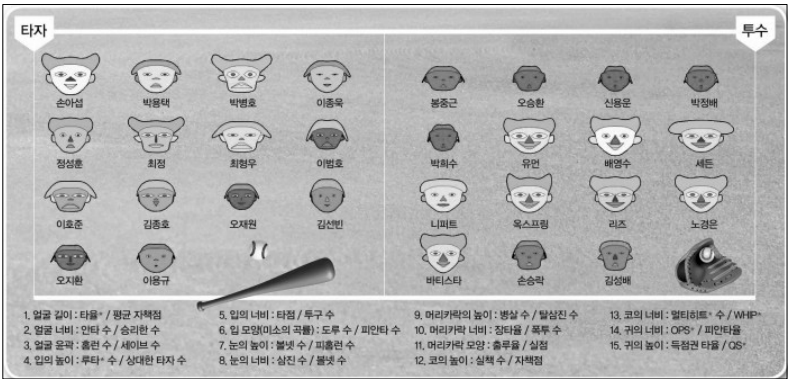
우리나라에서도 ≪월간동아≫에서 2013년 프로야구에서 활약하고 있는 선수들의 기록을 타자와 투수로 나누어 체르노프 얼굴로 아래와 같이 표현했다(조가현, 2013). 이 그림을 보면 타자의 경우는 병살수나 삼진수, 실책수와 같은 부정적인 변수가 전체의 20%이지만 투수의 경우는 50%가 넘어서 직접적인 얼굴의 비교는 의미가 적지만 비슷한 성적을 보이는 선수들을 구별한다는 점에서는 의미가 있다.

<그림 5> 미국인의 삶의 양상에 따른 체르노프 얼굴형태 표현



자료: Spinelli and Zhou, 2004.

<그림 6> 우리나라 프로야구 선수들의 통계에 따른 체르노프 얼굴



자료: 조가현, 2013.

(3) 체르노프 얼굴의 장단점과 공간적 함의

체르노프 얼굴에 대한 의미와 활용을 살펴봄으로써 체르노프 얼굴이 가지는 장점과 단점을 정리해 보고 이에 대한 함의를 공간적 측면에서 찾아보고자 한다. 이를 통해서 체르노프 얼굴이라는 기법으로 데이터를 지역이나 공간에 표현하는 방식이 다른 통계 기법과 다른 점이 무엇이고 공간적 인지효과는 무엇인지를 살펴본다.

우선 체르노프 얼굴을 통해서 데이터를 표현할 때의 장점은 체르노프

얼굴이 사람의 얼굴 요소들을 표현하고자 하는 데이터와 매칭해서 다차원적으로 표현하는 방식이기 때문에 다변량 자료 표현에 적합하다는 점이다. 이는 공간적으로 다양한 통계 값을 한꺼번에 표현할 때 유리하다. 지역이나 도시의 다양한 통계나 지표를 통합적으로 또는 한 가지로 표현할 수 있다는 뜻이다. 둘째, 체르노프 얼굴은 얼굴이라는 표현방식을 통해서 한 도시나 지역의 특징들을 쉽게 비교할 수 있다는 점이다. 체르노프 얼굴의 척도를 잘 활용한다면 특별히 순위(혹은 서열)척도의 데이터에 대한 비교를 쉽게 할 수 있고 이를 통해서 도시나 지역의 강점과 약점을 지역주민들이 쉽게 인식할 수 있다는 장점을 가진다.

반면에 체르노프 얼굴이 가지는 약점도 있는데 무엇보다도 데이터가 사람의 얼굴에 대입됨으로써 표정이 만들어지기 때문에 데이터나 지도에서 말하고자 하는 내용과는 별개로 감정이나 느낌을 표현할 수 있다는 점이다. 따라서 체르노프 얼굴의 각 요소들(머리 크기, 눈 크기 등)에 대한 명확하고 확실한 선정논리가 없다면 데이터를 왜곡할 가능성이 높아진다. 두 번째 약점은 얼굴의 각 부위와 사용되는 데이터와의 객관적인 매칭 기준을 세우기가 어려운 경우도 있다는 점이다. 따라서 체르노프 얼굴에서 주어진 데이터와 얼굴의 각 부위를 완벽하게 객관적으로 매칭해서 표현하는 것은 쉽지 않고 어느 정도 주관적인 기준이 들어가게 된다는 점이 약점으로 지적될 수 있다.

3. 분석내용과 결과

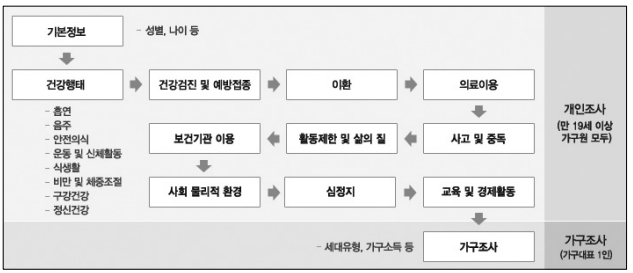
1) 지역사회건강통계

보건복지부 산하 질병관리본부에서는 지역사회의 건강 수준을 평가하기 위해서 2008년부터 매년 전국 만 19세 이상 성인을 대상으로 한 표본조사를 통해서 지역사회의 건강상태에 대한 통계(「지역사회건강통계」)를

발표해왔다.[3] 과거의 지역에 관한 보건이나 의료에 대한 통계조사는 지역 내의 병원병상의 수나 의사의 수와 같은 의료시설이나 의료인력으로 표현되는 지표를 활용하는 것이 보편적이었다. 그러나, 「지역사회건강통계」에서는 흡연, 음주, 안전, 운동, 식생활, 비만, 구강, 정신건강, 검진, 질환이환 등의 지역주민의 건강에 관한 생활양식에 대한 통계를 통해 지역의 건강에 대한 평가 및 진단을 한다는 데 그 차별적인 의미를 가진다고 볼 수 있다(<그림 7>과 <그림 8> 참조).

2015년 4월 7일에 발표한 『2014년 지역사회건강조사 결과』에 의하면 흡연, 음주, 운동, 비만 등의 지표들을 통해서 나타난 지역주민들의 건강상태가 여전히 미흡한 것으로 나타났다. 특히 고혈압, 당뇨병 등 만성질환자들은 건강 실천을 원하지만 아직 구체적인 실천이 쉽지 않은 것으로

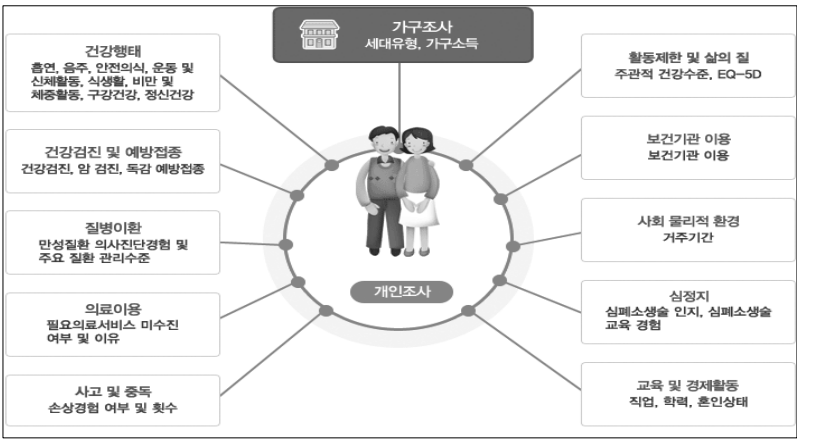
<그림 7> 지역사회건강통계 조사내용



자료: 질병안전본부 지역사회건강조사 홈페이지(https://chs.cdc.go.kr).

3) 지역사회건강통계의 조사개요는 아래와 같다(질병관리본부, 2015).
- 조사배경: 지역보건법(제4조 지역보건의료계획의 내용)에 따른 지역주민 건강수준의 모니터링과 평가.
- 조사목적: 지역보건의료계획 수립에 필요한 시, 군, 구 단위 건강통계 산출.
- 조사기간: 2014.8.16~2014.10.31(2.5개월) ※2008년부터 동일시기에 동일 조사체계로 매년 실시.
- 조사대상: 기초자치단체별(254개 보건소) 평균 900명(만 19세 이상 가구원 전원, 오차범위 ±3%), 총 227,700명.
- 조사방법: 가구방문, 전자조사표(CAPI) 1 : 1 면접조사.
- 조사문항 및 산출지표: 총 18개 영역, 168개 조사문항, 127개 산출지표.

<그림 8> 지역사회건강통계 조사영역



자료: 질병안전본부 지역사회건강조사 홈페이지(https://chs.cdc.go.kr).

보인다. 이번 조사에서 볼 수 있는 몇 가지 흥미로운 사실이 있는데, 정부의 정책이 국민의 건강에 영향을 줄 수 있다는 점이다. 예를 들면, 금연조례 시행 지역(36개월 이상)이 미시행 지역에 비해 흡연율 감소 경향을 보이는 것은 정부, 특히 지방정부의 건강과 관련된 정책에 효과가 있음을 보여준다. 더 나아가, 지난 7년간 지역의 주요 건강행태(남자 현재 흡연율, 고위험음주율, 걷기 실천율)에서는 지역적 격차가 뚜렷하게 나타나는데, 이러한 결과 역시 개인의 역할과 함께 지방정부나 지역사회의 노력이 개인의 건강과 무관하지 않다는 것을 간접적으로 보여준다.

2) 지역사회건강통계에 대한 광역자치단체별 체르노프 얼굴과 평가

본 연구에서는 체르노프 얼굴(Chernoff Face)을 활용해서 광역자치단체들의 지역사회건강통계를 표현하고자 한다.[4] 여기에서 다루고자 하는

4) 빅데이터는 그 정의에 따르면 규모, 다양성, 속도와 같이 기존의 데이터와는 물리적 성질이나 특성이 다른 것을 의미한다. 그러나 기존의 데이터도 표현기법을

<표 2> 체르노프 얼굴에서 표현하는 지역사회건강 항목과 설명

체르노프 얼굴 부위		지역사회 건강항목	의미	긍정적/부정적 지표
머리	머리높이	흡연	(남자 현재흡연율) 평생 5갑(100개비) 이상 흡연한 사람 중에서 현재 흡연자('매일 피움' 또는 '가끔 피움')의 비율	부정적 지표
	머리너비	음주	(고위험 음주율) 최근 1년 동안 한 번의 술자리에서 남자는 7잔(여자의 경우 5잔) 이상을 주 2회 이상 마신다고 응답한 사람의 비율	부정적 지표
눈	눈 높이	신체활동	(걷기 실천율) 최근 1주일 동안 1일 30분 이상 걷기를 주 5일 이상 실천한 사람의 비율	긍정적 지표
	눈 너비			
코	코 높이	비만	(체중조절 시도율) 최근 1년 동안 체중을 '줄이거나' 또는 '유지'하려고 노력했던 사람의 비율	부정적 지표
	코 너비			
입	입 높이	구강건강	(점심식사 후 칫솔질 실천율) 점심식사 후 칫솔질 한 사람의 비율	긍정적 지표
	입 너비			
귀	귀 높이	정신건강	(우울감 경험률) 최근 1년 동안 연속적으로 2주(14일) 이상 일상생활에 지장이 있을 정도의 슬픔이나 절망감을 경험한 사람의 비율	긍정적 지표
	귀 너비			
얼굴	얼굴길이	질병-1	(고혈압 진단자 비율) 30세 이상 조사응답자 중 의사에게 고혈압을 진단 받은 사람의 비율	부정적 지표
	얼굴너비	질병-2	(당뇨병 진단자 비율) 30세 이상 조사응답자 중 의사에게 당뇨병을 진단 받은 사람의 비율	부정적 지표

광역자치단체의 통계는 여덟 가지고 여기에는 ① 흡연(남자 현재흡연율), ② 음주(고위험음주율), ③ 신체활동(걷기 실천율), ④ 비만(체중조절시도율), ⑤ 구강건강(점심식사 후 칫솔질 실천율), ⑥ 정신건강(우울감 경험률), ⑦ 질병-1(고혈압 진단자 비율), ⑧ 질병-2(당뇨병 진단자 비율)가 포함된다. 각 통계자료는 질병안전본부에서 제공하는 지역사회건강조사의 자료를 통해

다르게 하면 빅데이터의 영역에 포함된다. 본 연구의 <그림 1>과 <표 1>에서 표현방식이 빅데이터의 처리과정에 포함되는 것을 보면 이러한 점을 이해할 수 있다. 본 연구에서 지역사회건강에 관한 여덟 가지 지표 자료를 기존의 표현방식과 다르게 체르노프 얼굴을 통해서 다차원적으로 표현했다는 점에서도 빅데이터로서 가치를 가질 수 있다.

서 구축했는데 가장 최근 것으로 제공되는 2014년 기준의 통계자료를 사용해서 분석했다.

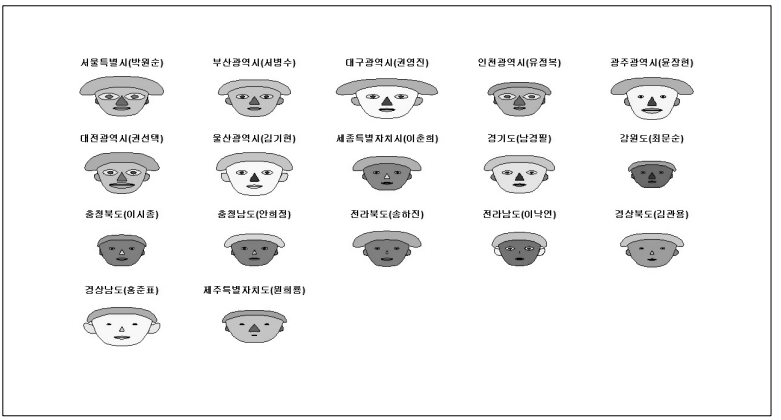
본 연구에는 광역지자체 원시자료에서 추출된 자료를 사용했다. 원시자료는 보건복지부 산하 질병관리본부에서 콜센터를 운영해 조사시점에 표본가구에 거주하는 만 19세 이상의 성인(1995년 7월 31일 이전 출생자)을 조사했고, 조사단위는 가구(가구조사) 및 가구원(개인조사)이었다(질병관리본부, 2015). 조사대상은 기초자치단체별(254개 보건소) 평균 900명(만 19세 이상 가구원 전원, 오차범위 ±3%)으로 총 22만 7700명이었다. 사실상 원시자료는 규모가 큰 빅데이터이기 때문에 본 연구에서는 원시자료의 규모나 이를 활용한 데이터 표현기법 모두 빅데이터 분석의 영역이라고 할 수 있다.

체르노프 얼굴에서 여덟 가지 통계가 의미하는 바는 <표 2>와 같다.[5] 이들 중에서 흡연(남자 현재흡연율), 음주(고위험 음주율), 정신건강(우울감 경험률), 질병-1(고혈압 진단자 비율), 질병-2(당뇨병 진단자 비율)는 부정적 지표라서 각 통계의 역수를 사용했으며, 신체활동(걷기 실천율), 비만(체중조절 시도율), 구강건강(점심식사 후 칫솔질 실천율)은 긍정적 지표라서 그 값을 그대로 활용했다.

체르노프 얼굴에서는 상대적 비교를 하기 때문에 얼굴의 각 부분이 클수록 더 좋은 지역사회 건강지표를 가진 자치단체인 반면에 작은 얼굴 요소를 가진 경우는 지역사회 건강지표가 나쁜 것을 나타난다. 다른 지표는 남성과 여성을 포함한 전체통계를 활용했지만 흡연율의 경우는 남성이 여성에 비해서 10배 이상 차이가 나기 때문에 남성 현재흡연율을

5) Turner(1979)의 연구에서 지역의 부유함(얼굴 모양), 고용률(입 모양), 도시생활 스트레스(입 모양), 백인인구비율(얼굴 색깔)을 표현한 것에도 어느 정도는 저자의 주관적인 기준이 들어갔다. 따라서 체르노프 얼굴을 완벽하게 객관적으로 표현하는 것은 쉽지 않고, 어느 정도 주관적인 기준이 들어가게 된다. 따라서 본 논문에서 제시한 여덟 가지 기준의 경우에도 입으로 표현되는 구강건강과 같이 연계성이 높은 경우도 있지만 귀나 코와 같이 연계성이 분명하지 않은 경우도 있게 된다.

<그림 10> 광역자치단체별 지역사회건강의 체르노프 얼굴



사용했다. 광역자치단체별 체르노프 얼굴의 표현은 <그림 10>에 나타나 있다.

(1) 흡연

체르노프 얼굴에서 머리의 높이로 표현되는 흡연율은 '남자 현재흡연율'을 기준으로 삼았으며 평생 5갑(100개비) 이상 흡연한 사람 중에서 현재 흡연자('매일 피움' 또는 '가끔 피움')의 비율로 나타나고 있다. 흡연율은 남성이 여성에 비해서 열 배 이상 높기 때문에 남성의 현재 흡연율을 사용하는 것이 더 보편적인 경향을 볼 수 있다고 판단된다.

높은 머리 모양을 한 서울특별시의 남성 흡연율이 가장 높고, 그 뒤를 전라북도, 대전광역시, 전라남도가 뒤따른다. 이러한 자치단체들은 상대적으로 다른 자치단체에 비해서 남성 흡연율이 낮은 지역이다. 생활 속에서 담배 냄새를 맡을 확률이 다른 자치단체에 비해서 다소 낮다고 할 수 있다. 반면에 낮은 머리모양의 인천광역시는 남성흡연율이 가장 높고, 그 뒤를 강원도와 제주특별자치도, 충청북도가 따른다. 관광중심지인 강원도와 제주특별자치도의 남성 흡연율이 높게 나타난 점이 흥미롭다.

(2) 음주

체르노프 얼굴에서 머리의 너비로 표현되는 음주의 경우는 '고위험 음주율'을 기준으로 삼았으며, 최근 1년 동안 한 번의 술자리에서 남자는 7잔(여자의 경우 5잔) 이상을 주 2회 이상 마신다고 응답한 사람의 비율을 의미한다.

넓은 머리모양을 보이는 대구광역시는 고위험 음주율이 가장 낮은 자치단체이고 그 뒤를 광주광역시와 서울특별시가 따르고 있다. 반면에 좁은 머리모양의 충청북도는 고위험 음주율이 가장 높은 자치단체다. 그 뒤를 제주특별자치도와 강원도가 따르고 있다. 제주특별자치도와 강원도의 주민들은 다른 지역에 비해서 음주와 흡연을 많이 하는 것으로 나타난다.

(3) 신체활동

체르노프 얼굴에서 눈의 높이와 너비로 표현되는 신체활동은 '걷기 실천율'을 기준으로 삼았으며 최근 1주일 동안 1일 30분 이상 걷기를 주 5일 이상 실천한 사람의 비율을 의미한다.

큰 눈의 높이와 너비를 보이는 서울특별시가 걷기 실천과 같은 신체활동을 가장 잘하고 있고, 그 뒤를 인천광역시와 대전광역시가 따른다. 반면에 작은 눈의 높이와 너비를 보이는 경상남도는 걷기 실천과 같은 신체활동을 잘하지 못하고 있으며 그 뒤를 제주특별자치도와 경상남도가 따른다. 영남권과 제주권 주민들의 신체운동 활성화를 위한 다양한 정책적 접근이 필요한 것 같다.

(4) 비만

체르노프 얼굴에서 코의 높이와 너비로 표현되는 비만은 '체중조절 시도율'을 기준으로 삼았으며 최근 1년 동안 체중을 '줄이거나' '유지'하려고 노력했던 사람의 비율을 의미한다.

큰 코의 모양을 보이는 자치단체들은 인천광역시, 서울특별시, 대전광

역시와 같은 대도시들이 차지하고 있는 반면에, 작은 코의 모양을 보이는 자치단체들은 전라남도, 전라북도, 경상북도, 경상남도와 같은 도 단위의 자치단체들이다. 즉 시 단위의 자치단체들에서 주민들은 비만에 대해서 인식을 하고 이를 관리하기 위해서 노력을 많이 하는 반면에 도 단위의 자치단체들에서 이러한 인식과 노력은 상대적으로 부족하다고 할 수 있다.

(5) 구강건강

체르노프 얼굴에서 입의 높이와 너비로 표현되는 구강건강은 '점심식사 후 칫솔질 실천율'을 기준으로 삼았는데 이는 조사대상자들 중에서 점심식사 후 칫솔질 한 사람의 비율을 의미한다.

큰 입 모양을 보이는 대전광역시, 광주광역시, 서울특별시는 구강건강에 대한 주민들의 관심이 작은 입 모양을 보이는 제주특별자치도, 경상북도, 강원도 주민들에 비해서 높다고 할 수 있다.

(6) 정신건강

체르노프 얼굴에서 귀의 높이와 너비로 표현되는 정신건강은 '우울감 경험률'을 기준으로 했는데, 이는 최근 1년 동안 연속적으로 2주(14일) 이상 일상생활에 지장이 있을 정도의 슬픔이나 절망감을 경험한 사람의 비율을 의미한다.

큰 귀 모양을 보이는 경상남도가 우울감을 경험한 주민의 비율이 가장 낮았고 그 뒤를 전라남도와 울산광역시가 따르고 있다. 반면에 작은 귀의 모양을 하고 있는 충청북도는 우울감을 경험한 주민의 비율이 가장 높았는데 경상남도에 비해서 2배 정도 많은 주민들이 우울감을 경험한 것으로 보인다. 충청북도의 뒤를 이어서 인천광역시, 강원도, 서울특별시가 그 뒤를 따르고 있다. 정신건강에 대한 지표를 인구구성 등에 대해 좀 더 심도 있게 분석함으로써 위와 같이 나타난 이유를 밝혀낼 수 있을 것이다.

(7) 질병-1(고혈압)

체르노프 얼굴에서 얼굴의 길이로 표현되는 질병-1은 '고혈압 진단자 비율'인데 30세 이상 조사응답자 중 의사에게 고혈압을 진단 받은 사람의 비율을 의미한다.

상대적으로 가장 긴 얼굴을 하고 있는 울산광역시에서 고혈압 진단자의 비율이 가장 낮았고, 그 뒤를 광주광역시, 대구광역시, 경상남도가 따르고 있다. 얼굴의 길이가 짧은 강원도, 충청남도, 전라남도, 충청북도에서는 고혈압 진단자의 비율이 높은 것으로 나타난다.

(8) 질병-2(당뇨병)

체르노프 얼굴에서 얼굴의 너비로 표현되는 질병-2는 '당뇨병 진단자 비율'인데 30세 이상 조사응답자 중 의사에게 당뇨병을 진단 받은 사람의 비율을 의미한다.

상대적으로 가장 넓은 얼굴을 한 대구광역시에서 당뇨병 진단자의 비율이 가장 낮았고, 그 뒤를 경상남도와 제주특별자치도가 따르고 있다. 얼굴의 길이가 좁은 강원도, 전라남도, 전라북도에서는 당뇨병 진단자의 비율이 높은 것으로 나타난다.

(9) 종합평가

종합적으로 보았을 때 특별시와 광역시에서의 지표들이 우수하게 나타나는 것을 얼굴 모양에서 알 수 있다. 대전광역시, 서울특별시, 대구광역시가 우수한 편이다. 반면에 도 단위 자치단체들의 성적은 다소 저조한데, 특히 강원도, 충청북도, 충청남도, 제주특별자치도에서는 개선을 위한 노력이 필요한 것으로 판단된다.

4. 결론

체르노프 얼굴(Chernoff Face)은 다차원 통계 데이터를 사람의 얼굴 이미지를 이용해 시각적으로 표현하는 것으로 최근 들어 빅데이터 분석의 표현기법으로 활용된다.

본 연구에서는 체르노프 얼굴을 활용해 우리나라 광역자치단체들의 지역사회건강에 대한 국민체감지표를 살펴보았다. 여기에서 다루고자 하는 광역자치단체의 통계는 여덟 가지로 ① 흡연(남자 현재흡연율), ② 음주(고위험음주율), ③ 신체활동(걷기 실천율), ④ 비만(체중조절시도율), ⑤ 구강건강(점심식사 후 칫솔질 실천율), ⑥ 정신건강(우울감 경험률), ⑦ 질병-1(고혈압 진단자 비율), ⑧ 질병-2(당뇨병 진단자 비율)가 포함된다. 각 통계자료의 경우는 질병안전본부에서 제공하는 2014년도 기준 지역사회건강조사의 자료를 통해서 구축했다.

분석결과를 종합적으로 보았을 때 특별시와 광역시에서의 지표들이 우수하게 나타나는 것을 얼굴 모양에서 알 수 있다. 대전광역시, 서울특별시, 대구광역시가 우수한 편이다. 반면에 도 단위 자치단체들의 성적은 다소 저조한데, 특히 강원도, 충청북도, 충청남도, 제주특별자치도에서는 개선을 위한 노력이 필요한 것으로 판단된다. 이는 구체적으로는 지역주민의 건강증진을 위한 다양한 홍보 및 캠페인, 체육시설 신설, 확충 및 관리, 그리고 자치단체 내의 보건소 및 병원 등을 통한 건강증진 프로그램 개설 등으로 나타날 수 있다. 이와 함께 지역주민들의 건강한 삶에 영향을 줄 수 있는 녹지지표나 환경지표들에 대한 개선을 위한 다양한 전략을 세우는 것도 효과적일 것이다.

빅데이터 분석 및 표현기법의 하나인 체르노프 얼굴을 활용한 지역분석과 표현기법은 기존의 데이터분석이나 표현을 향상시키기 위한 좋은 사례가 될 수 있다. 이를 지역홍보 및 마케팅에 활용할 수 있고, 지역 주민들이 쉽게 이해할 수 있는 얼굴 형식이라는 점에서 선거전략으로도 발전시킬 수 있는 가능성이 있다. 그렇지만 이 방식에는 몇 가지 약점도

나타난다. 데이터가 사람의 얼굴에 대입됨으로써 표정이 만들어지기 때문에 데이터나 지도에서 말하고자 하는 내용과는 별개로 감정이나 느낌을 일으킬 수 있다는 점이나 얼굴의 각 부위가 어떤 변수를 표현할지에 대한 객관적인 기준이 부족하다는 점 등은 극복되어야 할 한계로 남아 있다.

원고접수일: 2015년 12월 17일
심사완료일: 2016년 1월 18일
게재확정일: 2016년 2월 2일
최종원고 접수일: 2016년 2월 3일

Abstract

A Big Data Analysis of Urban Statistics Expression –Chernoff Face-Based Expression of Local Community Health Index in Korea

Ki, Jung-hoon

Chernoff face is a statistical visual expression method of multivariate data using the human face image. It matches human face' elements such as eyes, nose, mouth to each variable of multivariate database. The method's major advantage is that it is able to demonstrate various urban and regional statistics in one face and give the spatial simplicity of perception. Thus it is widely applied to big data analysis.

This study shows Korea metropolitan governments' local community health index using the Chernoff face. The health index includes ① smoking habit, ② alcohol drinking habit, ③ walking habit, ④ obesity, ⑤ teeth health, ⑥ mental health, ⑦ hypertension, and ⑧ diabetes. Data source is the 2014 Korea Local Community Health Survey that the Korea Centers for Disease Control and Prevention provides. Generally speaking, metropolitan areas such as Daejeon, Seoul, and Daegu shows better performance in the health index than provincial areas such as Kangwon, Chungcheongbuk, Chungcheongnam, and Jeju which need a policy measure to improve people's health habits.

Although the Chernoff face can be a better research tool or marketing method for local governments and elections, it has some demerits such as lack of clear standard for matching the facial elements and statistical variables.

Keywords: Big Data Analysis, Chernoff Face, Local Community Health Index

참고문헌

기정훈. 2014a.『인문사회과학분야에서의 빅데이터 활용과 실제』. 명지대학교 빅데이터 분석 연구소 정기세미나. 2014년 10월 16일. 명지대학교 빅데이터 분석 연구소.

_____. 2014b.『빅데이터를 활용한 사회과학 연구동향 분석 및 적용가능성』. 제4차 콜로키움. 2014년 11월 11일. 명지대학교 사회과학연구소.

송길영. 2011.『Understanding Society through SOCIALmetrics』.

정지선. 2012.『성공적인 빅데이터 활용을 위한 3대 요소』. 한국정보화진흥원.

조가현. 2013.「통계데이터를 얼굴 모양으로 나타내는 체르노프 얼굴」. ≪월간동아≫.

질병관리본부. 2015.『2008~2014 지역사회건강조사 조사개요 및 주요결과』. 보건복지부.

한국콘텐츠진흥원. 2014.『빅데이터 시장현황과 콘텐츠산업분야에 대한 시사점』. 한국콘텐츠진흥원.

Chernoff, H. 1973. "The Use of Faces to Represent Points in K-Dimensional Space Graphically." *Journal of the American Statistical Association*, Vol. 68, No. 342, pp. 361~368.

Huffman, D. 2010. "on the abuse of chernoff faces". Cartastrophe. http://cartastrophe.wordpress.com/2010/06/16/on-the-abuse-of-chernoff-faces

Lee, J. et al. 2013. "Good Bank Evaluation by Chernoff Face Analysis using SAS macro faces." *The Korean Journal of Applied Statistics*, 26(6), pp. 959~975.

Spinelli, J. G. and Zhou, Y. 2004. "Mapping Quality of Life with Chernoff Faces." Unpublished paper.

Thoplan, R. 2014. "A Statistical Graphic Exploration of Crime Rate for the States of USA." *Research Journal of Social Science and Management*, 4(6), pp. 58~67.

Turner, E. 1979. *Life in Los Angeles 1970*. Los Angeles Community Analysis Bureau.

Warden, P. 2011. *Big Data Glossary*. O'Reilly Media.