한국문화관광연구원
Korea Culture & Tourism Institute

문화관광 인사이트

제 117 호

문화·관광 인사이트
제117호
2018. 05. 30
발행처-한국문화관광연구원
www.kcti.re.kr

빅데이터를 활용한 관광통계 생산방안

권태일 | 통계평가센터 부연구위원※

1 들어가며

디지털 경제의 확산은 데이터와 정보의 폭발적인 증가로 이어져 '빅데이터(BIg Data)' 시대가 도래 하였다. 데이터 기반(data-driven) 사회에서 대규모 데이터는 국가경쟁력을 좌우하는 중요한 '자원'이다. 미국과 영국은 빅데이터 시대를 맞이하여 공공 정보의 전면적인 개방과 데이터 활용을 통한 가치 창출을 국가 전략으로 내걸고 새로운 혁신을 도모하고 있다. 우리나라도 정부 차원에서 사회 현안에 대한 선제적 대응수단으로 빅데이터를 적극 활용하고 있으며, 정책 결정 과정에서 데이터의 역할이 강조될수록 통계의 중요성은 더욱 부각되고 있다.

21세기 새로운 자산으로 부각되고 있는 빅데이터를 관광분야에서 객관적이고 시의성 있게 활용할 수 있는 방안을 마련하는 것은 매우 시급한 일이다. 이에 본 연구에서는 관광 빅데이터의 원활한 활용을 위해 다음과 같은 내용을 검토하였다.

첫 번째, 빅데이터 중 현재 활용 가능한 통신 데이터의 현황을 검토하였으며 두 번째, 관광분야에서 빅데이터의 활용을 위한 설계방안을 제시하였다. 세 번째 설계방안을 기준으로 실제 통신데이터를 활용한 실증분석을 실시하였으며, 마지막으로 보다 나은 활용을 위한 빅데이터의 한계 및 문제점을 제시하였다. 본 연구 결과가 향후 관광분야 빅데이터 활용의 기초적 자료가 되길 기대해본다.

2 통신 빅데이터 개념 및 설계

■ 통신 빅데이터 현황

본 연구에서 활용된 S사의 통신 빅데이터는 하루 약 280테라바이트(Terabyte)의 데이터가 생성되고 있으며, 이러한 통신 트래픽 데이터는 하둡을 통한 추출·정제 및 수집·저장을 위한 프로세스 체계를 구축하여 분석에 활용되고 있다. S사 통신 빅데이터는 무선 통신 방식에 따른 1X, 2G, 3G, 4G, LTE 등의 모든 사용자의 CDR(Action 기반) 데이터와 Signal(Location 기반) 데이터를 적용하고 있다. 집계방식은 사용 목적에

※ tikwon@kcti.re.kr, 02-2669-8902
이 글은 2017년 수시과제로 수행된 권태일·이충희, 「관광분야 빅데이터 분석 및 활용체계 연구」(한국문화관광연구원) 결과를 토대로 작성하였음.

따라 내국인의 ① 인구유입량 데이터[1] ② OD 동선 데이터[2]로 구분 할 수 있으며, 외국인의 경우 로밍데이터를 활용하여 국적별 유입량 데이터와 OD 동선 데이터의 수집 및 확보가 가능하다.

내국인에 대한 OD 동선 데이터는 하루에 약 6억 6천만건의 원시자료가 생성되고 있으며 2016년 10월 이후부터 데이터를 비식별화하고 DB로 구축하여 분석에 활용하고 있다. 본 연구에서는 인구통계학적으로 정의된 용어들을 중심으로 살펴본 후, 관광분야에서의 통신 데이터에 대한 개념을 검토하였다.

표-1 | 통신데이터의 개념

분류	시간별 유니크 데이터	일별 유니크 데이터	실시간 데이터	개인화 데이터
조건	일별 시간대 중복 데이터	일별 데이터	일별 15분 단위 데이터	1시간 개인화 데이터
기간	2013년 5월부터 축척	2014년 1월부터 축척	2016년 1월부터 축적	최근 3개월
단위	1시간 단위	일 단위	15분 단위	1시간 단위
형태	통계청 집계 데이터	통계청 집계 데이터	통계청 집계 데이터	개인화 데이터
유형	활동인구	존재인구	존재인구	개인별 위치 데이터
내용	유동인구 데이터는 1시간 단위로 중복된 인구를 산정하여 좁은 영역에 활동하는 인구가 많은 지역을 추출할 때 용이한 자료	유입인구 데이터는 일 단위로 시간단위 중복이 없는 인구를 산정하여 설정 영역에 하루에 존재했던 인구를 추출할 때 용이한 자료	실시간 데이터는 15분 간격으로 스냅샷으로 인구를 산정한 자료로서 서비스 인구 추정에 활용 15분 기지국 최종 목적지 정보를 포착하여 모바일 폰 사용자를 산정하는 데이터	개인화 데이터는 모바일 폰 사용자의 동선파악 및 체류시간을 위하여 개인의 개별데이터로서 개인정보 보호를 위하여 최근 3개월 자료만 분석이 가능한 데이터

■ 통신 빅데이터 설계

통신 빅데이터의 관광분야 활용[3]을 위해 6단계 프로세스를 기준으로 설계방안을 도출하였다.

① 통신데이터의 대표성 검증 : 특정 통신사가 가지고 있는 거주지별 가입자에 대한 대표성을 검증하기 위해 통계청 인구센서스 기준으로 전국의 성·연령별 분포를 파악하였다. 통신사의 가입자(청구지 정보)와 통계청 인구센서스의 매칭 결과 지역간 정확한 배분이 되지 않고 있음을 파악하였으며, 이에 3단계에 조건을 거쳐 통신 가입자의 실거주지를 추정하였다. 조건1은 청구지 200m 이내의 기지국과 50% 이상 통신한 경우, 조건2는 청구지 200m~500m 이내 기지국과 50% 이상 통신한 고객 중 해당 기지국이 가장 가깝고 청구지가 해당 기지국 pCell 내에 위치한 고객, 조건3은 조건1과 조건2에 해당하지 않고 특정 기지국과 50% 이상 통신한 가입자를 대상으로 실거주지를 파악하였다. 이를 통해 전체 가입자의 86.7%는 특정 기지국과 50% 이상 통신하는 실 거주지로서 추정이 가능한 샘플로 활용가능하다는 판단을 하였다.

② 데이터 집계를 위한 표본설계 : 기존 빅데이터에서는 OD개념의 이동량 생산통계는 없으며, 방문지 기준의 이동량 통계의 경우 통신3사의 가입율로 단순 보정하여 사용함으로써 통계 생산의 객관성 신뢰성의 확보가 어려웠다. 이에 관광분야 OD 통계 생산을 위해 S사 통신가입자 2,400만명 가운데 600만명의 정확한 거주지를 파악하고 패널화 하였으며, 2개의 표본설계안을 검토하였다. 1안은 통계청 집계구 기준 (10만개->2만개 계통추출), 2안은 230개 시·군·구 단위 기준이다. 최종적으로는 1안을 기준으로 실증분석을 실시하였다.

③ 관광OD 도출을 위한 기준 검토 : 통신데이터를 활용하여 관광OD를 산출하기 위해 관광객에 대한 정의를 실시하였다. 전체 이동데이터 가운데 거주자, 통근통학, 2시간이하 해당지역 체류자는 제외하였으며, 이외 모든 이동량 데이터를 기준으로 시·군·구간 관광목적비율을 적용하여 최종 OD데이터를 산출하였다. 국민여행실태조사의 경우 17개 광역시도를 기준으로 관광이동총량을 추정하였으나, 본 연구에서는 230개 기초단위의 모든 이동을 포함하여 추정함으로써 보다 세부적인 결과값 도출이 가능해졌다.

④ 관광 주요지표 도출 : 통신 빅데이터를 활용한 관광 주요지표 도출은 국민여행 실태조사의 대표 지표[4]를 기준으로 설정하였다.

⑤ 모수추정 : 모수는 통계청 집계구 가중치를 활

1) 인구 유입량 데이터의 추출 조건은 각 시·군·구 방문객 총 수를 나타내며, 경우에 따라 해당 시·군·구 인구와 타 지역에서 유입되는 인구를 구분할 수 있는 데이터임.

2) OD 동선 데이터는 해당 시·군·구의 야간 체류자를 거주인구(Origin)로 가정하고, 2,700만 통신 폰 사용자가 전국의 시·군·구 중 거주지역이 아닌 타 시·군·구에 2시간 이상 체류한 중간 지역을 경유지로 하며, 귀가하기 전의 최종 시·군·구를 목적지(Destination)로 설정하여 추출한 데이터임.

3) 관광분야 활용은 빅데이터를 활용한 시의성 있는 내·외국인의 이동량 추정임.

4) 대표지표 : 여행횟수, 여행일수, 이동총량 등

용하여 최종 도출하였으며, 도출된 데이터를 바탕으로 당일은 1명, 숙박은 평균숙박일수를 기준으로 나누어서 1명으로 보정하고 해당 월에 거주지별 순수 방문객 수를 파악하였으며, 체제일수를 고려한 관광 이동총량 또한 파악하였다.

⑥ 데이터 검증 : 마지막으로 기존에 생산되고 있는 비교 가능한 실측값(제주도 입도객 통계)을 기준으로 도출된 관광분야 빅데이터 결과값과 비교 검증을 통해 데이터 활용의 정확성을 검토하였다.

3 통신 빅데이터 실증분석

본 연구에서는 앞서 설계된 빅데이터 활용방법을 바탕으로 국·내외 관광객의 OD를 파악하기 위해 다양한 연구를 시도하였다. 첫 번째, 데이터의 객관성 및 정확성을 확보하기 위해 통신 빅데이터와 제주도 입도객 통계를 일(日)단위로 비교하였으며 두 번째 검증된 데이터를 바탕으로 2017년 7월 기준 관광객의 OD별 이동총량(숙박+당일)을 분석하였다. 세 번째 이동총량 데이터를 기준으로 숙박객의 중복을 제거하고 지역별 관광인구(숙박+당일)를 도출하였다. 네 번째 외국인 로밍데이터를 바탕으로 국적별×시도별 OD간 이동량을 도출하였다. 해당 연구는 통신 빅데이터를 기반으로 거주지(Origin)에서 방문지(Destination)[5]로의 이동을 추정하였으며 관광목적의 데이터를 도출하기 위해 일상적 활동인 통근통학을 제외한 통신인구 이동 중 교통연구원에서 제공하는 시·군·구간 이동목적별 비율에서 업무, 여가오락친지방문, 쇼핑에 대한 목적별 비율을 반영하여 최종 결과값을 산출하였다.

■ 통신 빅데이터와 제주도 입도객 통계 비교

통신 빅데이터의 시·도간 OD 이동량을 파악하기 위해서는 해당 데이터에 대한 정확성 검증 작업이 필요하다. 해당 데이터 사용의 당위성을 검증하기 위해 본 연구에서는 제주도의 입도객 통계와 통신 빅데이터의 일간 비교를 통해 오차의 차이를 검증하였다. 제주도를 방문하는 유입인구의 평균 체류일수는 3.31일이며 [표 2]와 같이 나타났다. S사의 통신 데이터의 트래픽으로 파악하는 유입인구는 현재 제주도에 온 방문객과 또는 숙박을 통하여 체류하는 유입인구 모두가 포함되어 있다. 또한 입도객 통계에서 모바일폰 미소지자가 포함되어 있음을 볼 때 신뢰할 만한 결과로 말할 수 있다(약 97.5%의 커버리지).

표-2 | 통신데이터와 제주도 입도객 통계 비교(2017년 7월, 명)

일자	통신데이터 통계(a)	제주도입도객 통계(b)	데이터 차이 c=b-a
20170701	120,143	122,351	2,208
20170702	119,332	123,311	3,979
20170703	111,032	118,213	7,182
20170704	104,693	98,966	(5,727)
20170705	115,997	115,310	(687)
20170706	115,671	125,062	9,391
20170707	128,110	124,453	(3,657)
20170708	136,851	120,544	(16,307)
20170709	131,793	121,689	(10,104)
20170710	119,130	125,211	6,081
20170711	116,465	123,218	6,753
20170712	125,062	127,299	2,237
20170713	130,281	127,693	(2,588)
20170714	129,078	120,732	(8,346)
20170715	131,403	127,753	(3,650)
20170716	128,901	131,563	2,662
20170717	121,262	127,121	5,858
20170718	116,676	116,783	107
20170719	113,604	117,793	4,188
20170720	114,052	127,998	13,945
20170721	114,678	121,593	6,915
20170722	119,600	131,116	11,516
20170723	120,586	130,457	9,871
20170724	115,192	124,019	8,828
20170725	113,573	115,191	1,618
20170726	116,966	133,784	16,818
20170727	122,728	133,542	10,814
20170728	132,902	144,839	11,936
20170729	146,641	158,450	11,809
20170730	152,931	152,505	(426)
20170731	153,045	146,276	(6,770)
평균	123,819	126,930	3,111

■ 국내여행 총량 산출

본 연구에서는 실증분석을 위해 통신 빅데이터를 활용한 국내여행 참가자수, 참가횟수, 이동총량에 대한 숙박여행과 당일여행에 대한 결과값을 도출하였다. 국

5) 외국인의 경우 국가와 방문지간의 비교 실시.

내여행 총량은 통신 빅데이터 활용에 대한 설계 내용을 바탕으로 지역간 OD를 도출한 후 통근통학 인구 제외(거주지를 기준으로 주중에 주·야간(07-09시, 19시-21시) 일정한 패턴으로 움직이는 인구)하고 이를 바탕으로 교통연구원의 시·군·구간 관광목적별 이동비율을 적용하여 산정하였다. 관광목적별 이동비율은 업무, 여가오락친지방문, 쇼핑에 해당되는 비율(60%)을 적용하였다. 분석결과 아래 [그림1]과 같이 경기도 지역의 이동량이 가장 높게 나타났으며, 세종지역의 이동량이 가장 낮게 나타났다.

외국인 관광객의 경우, 외래관광객실태조사와 동일비교를 위해 통신사 로밍데이터를 기준으로 국가별×지역별 OD 통계를 생산하였다. 외국인 이동총량의 경우 법무부의 월별×국가별 입국자 수를 기준으로 가중치를 통한 최종값을 도출하였다. 최종 결과는 [그림2]와 같이 중국이 가장 높은 방문현황을 보이고 있으며 일본, 대만, 미국, 홍콩 순으로 외국인 방문객 수가 높게 나타났다. 또한 외국인 방문객의 이동량은 국적과 지역별로 구분하여 분석결과의 도출이 가능하기 때문에 향후 보다 활용도가 클 것으로 판단된다.

그림-1 | 내국인의 지역별 이동량 추정(2017년 7월 기준)

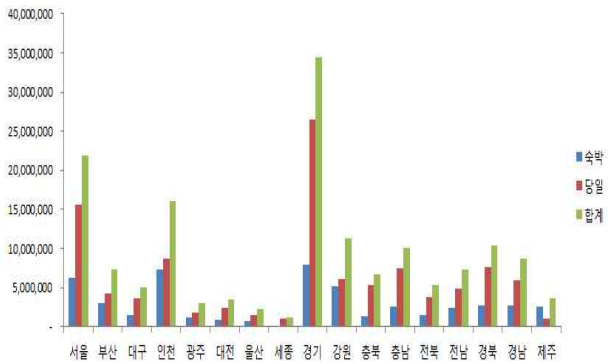
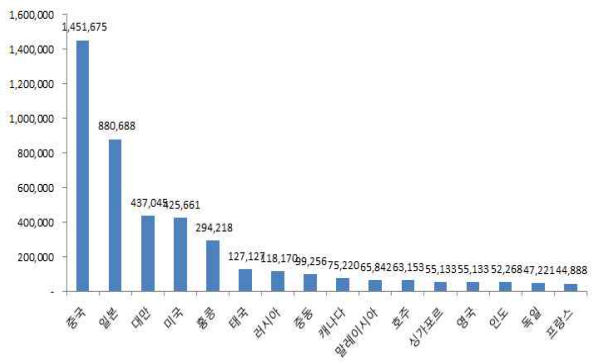


그림-2 | 외국인의 국가별 방문현황 추정



4 통신 데이터 한계 및 문제점

우리나라의 경우 이동통신 3사에 의하여 통신시장이 운영·관리되고 있다. 개념적으로 통신 3사의 모든 정보를 전수데이터(population data)로 하여 시간, 일, 주간, 순기, 월, 년 등 시공간 제약이 없는 분석데이터를 수집할 수 있는 것처럼 생각할 수 있겠으나, 기지국의 구조적 상이함 등으로 인한 기술적 통합은 불가능에 가까운 상황이며 제도적으로도 주도할만한 정부기관이 없고, 개인정보보호법 또는 민간기업의 한계 등으로 추진이 어려운 실정이다.

■ 내국인 통신 데이터 한계 및 문제점

통신 빅데이터는 인구의 이동 및 변화에 대하여 실시간적 결과를 보여준다. 이러한 인구는 관광/레저인 여행 형태이거나, 생활밀착형 통근·통학 등 여러 목적으로 이동한다. 관광을 광의의 개념으로 보면 친지방문에서 비즈니스 출장까지 거주지와 일터를 떠나는 것 일체를 의미한다[6]. 통신 빅데이터의 특성상 이동경로를 명확히 파악할 수는 있지만 이동의 목적을 확인할 수는 없다. 기존 분석에서는 이러한 한계를 극복하기 위하여 한국교통연구원(국토교통부산하)에서 실시하는 교통수요 분석 기초조사 결과를 바탕으로 여가/오락/친지방문의 여행 목적별 비율을 반영하여 전체 인구 유입량에서 관광을 목적으로 하는 인구를 산정하였다.

본 연구에서는 1차적으로 교통수요분석 결과를 바탕으로 시·군·구간 목적별 OD비율을 적용한 실증분석을 시도하였으나 향후, 보다 시의성 있고 객관적인 자료 도출을 위해서는 국민여행실태조사와의 연계를 통한 보다 세부적인 관광목적의 지역간 이동현황 파악이 필요할 것이다.

■ 외국인 통신 데이터 한계 및 문제점

외국인에 대한 로밍데이터는 S사의 점유율이 약 50%이다. 그러나 국가별로 국내 통신사와 계약된 비율이 다르고 외국인 핸드폰의 on/off 비율이 빈번히

6) 국민여행실태조사 여행 목적 : 여가/위락/휴가, 건강/치료, 종교/성지순례, 가족/친척/친구/친구 방문, 교육/훈련/연수, 쇼핑, 사업 및 전문활동/업무상 목적, 기타.

발생하기 때문에 정확한 실시간 트레픽을 잡아내는 데는 한계가 있다. 유럽과 같은 대륙은 1개의 메이저 통신사가 여러 국가의 이동통신 시장을 점유하고 있어 통신사 간 계약에 의한 로밍데이터 추출은 특정 대륙의 특정 국가의 국적을 구별하는데 있어 일부 오차를 발생시킨다.

또한 외국인 로밍데이터는 국적이나 인원을 파악할 수 있지만, 인구특성(성 및 연령그룹별)을 파악할 수는 없다. 더구나 외국이 방문객 보정을 위한 출입국통계는 월별 공표자료이지만, 잠정치와 확정치의 문제로 정확한 시간적인 검토가 필요하다. 향후 로밍데이터를 활용한 외국인 관광객의 빅데이터 결과분석을 위해서는 보다 다차원적인 검토가 필요할 것으로 판단된다.

5 향후 빅데이터 활용방안

■ 빅데이터 활용한 예측시스템 개발

관광 빅데이터의 객관적이고 신뢰성 있는 생산이 가능해지면 과거 데이터의 축적과 함께 미래에 대한 예측이 가능한 데이터로 활용될 수 있다. 다만 빅데이터를 활용한 예측시스템을 구축하기 위해서는 관련 기반투자와 통합적으로 관리할 수 있는 조직이 전제되어야 할 것이다.

지금까지의 빅데이터 활용은 현황 중심의 성과 점검에 초점 맞춰져 있었다. 또한 명확한 기준이 없는 비슷한 사업의 혼재 속에서 빅데이터 신뢰성 문제 제기 등으로 온전하게 사용되지 못하고 있는 실정이다. 이에 본 연구에서는 관광 빅데이터 활용을 위한 명확한 개념과 추정 프로세스를 제시함으로써 보다 정교하고 신뢰성 있는 결과 도출 및 예측 가능한 기반을 마련하였다. 향후 보다 체계적 생산과 활용을 통해 관광의 성과를 진단하고 정책을 수립하기 위한 예측에도 빅데이터가 기여할 것으로 기대된다.

■ 통계작성승인제도 활용을 통한 검증 제안

한국문화관광연구원은 통계작성지정기관[7]이다. 본 연구원에서 작성하는 통계 중 국가승인통계의 규격과 조건을 충족하면, 통계작성승인제도의 절차에 따라 통계청으로부터 승인을 받을 수 있다. 이에, 본 연구의 분석 결과를 토대로 관광 빅데이터의 국가승인통계 승인 추진을 진행할 필요가 있다.

통계작성승인 제도의 취지는 첫째, 무분별한 통계 생산을 막아 유사·중복 통계작성을 사전에 예방하여 인력 및 예산 낭비요인을 제거함에 있다. 둘째, 통계의 기술적 측면에서 작성하고자 하는 내용 검토 후 예상되는 문제점을 사전에 제거하여 신뢰성 높은 통계를 생산함에 있다.[8]

현재까지 우리나라에서 빅데이터를 활용한 국가승인통계는 1회 한정적으로 승인 받은 경우를 제외하고는 없었다. 본 연구의 결과를 국가승인통계로 추진하기 위해, 기존의 빅데이터 국가승인통계 사례를 검토하고 관련 관계자들을 통해 유사·중복 문제에 대해 검토 한 바 있다. 아울러, 통계의 기술적 측면에서 예상되는 문제점 등에 대해 산·학·연과의 자문 등을 통해 논의하였다.

통계작성승인 제도가 통계 작성에 관련한 모든 문제를 해결 할 수 있는 전제조건은 아니다. 그러나 국가승인통계로 승인을 받기 위해서는 통계작성에 관한 신뢰성 확보, 전문성을 강화한 통계작성 체계 구축이 전제되어야 하기 때문에, 통계법에 따른 통계작성 및 관리체계의 도입은 정책적으로 필요한 부분이다. 이와 같은 판단 하에, 지속적으로 산학연의 전문가 자문 및 통계청과의 협의·조정을 통해 국가승인통계로 추진하기 위해 구체적인 논의를 진행해야 할 것이다.

본 연구결과가 향후 관광분야 빅데이터 활용의 기초자료가 되길 기대해 본다.

7) "통계작성기관"이란 중앙행정기관·지방자치단체 및 제15조에 따라 지정을 받은 통계작성지정기관을 말함(통계법 제3조제3호). 한국문화관광연구원은 2016년 11월 8일, 통계작성지정기관으로 지정 받음.
8) 통계청(2015),「통계조정업무 매뉴얼」, p.33